# O IMPÉRIO DA IA

Por dentro da corrida irresponsável pela dominação total

KAREN HAO

Rocco DIGITAL

KAREN HAO

# O IMPÉRIO DA IA

Por dentro da corrida irresponsável
pela dominação total

Tradução de Ananda Alves e André Sequeira

Rocco
DIGITAL

*Para minha família,*
*passada, presente e futura.*

*Aos movimentos ao redor do mundo*
*que se recusam à desapropriação*
*em nome da abundância.*

Dizem[1] que explicar é desfazer a explicação. Essa afirmação nunca se aplicou tão bem quanto na área de programação de computadores — principalmente no que é conhecido como programação heurística e inteligência artificial —, pois, em ambos os domínios, as máquinas são feitas para se comportarem de maneiras maravilhosas, frequentemente capazes de deslumbrar até o observador mais experiente. Entretanto, uma vez que um programa é exposto e seu funcionamento interno é explicado por meio de uma linguagem suficientemente clara para gerar compreensão, sua magia se desfaz, e ele se revela uma mera coleção de procedimentos, cada um bastante concebível. O observador então pensa: "Eu poderia ter criado isso." Com essa concepção, ele move o programa em questão da prateleira rotulada como "inteligente" para aquela destinada às curiosidades — adequadas apenas para serem discutidas com pessoas menos esclarecidas do que ele.

**— JOSEPH WEIZENBAUM, PROFESSOR DO MIT E CRIADOR DO ELIZA, O PRIMEIRO CHATBOT, EM 1966.**

"Pessoas bem-sucedidas[2] criam empresas. Pessoas mais bem-sucedidas criam países. As pessoas mais bem-sucedidas de todas criam religiões."

Ouvi isso de Qi Lu, mas não tenho certeza da origem. Essa frase me fez pensar… Os fundadores mais bem-sucedidos não começam com a intenção de criar empresas. Eles estão em uma missão para criar algo mais próximo de uma religião e, em dado momento, percebem que fundar uma empresa é simplesmente a maneira mais fácil de atingir seu objetivo.

**— SAM ALTMAN, 2013.**

# SUMÁRIO

Para pular o Sumário, clique aqui.

NOTA DA AUTORA
PRÓLOGO: Uma corrida pelo trono

I

1 – Direito divino
2 – Uma missão civilizadora
3 – Centro nervoso
4 – Sonhos de modernidade
5 – Escala de ambição

II

6 – Ascensão
7 – A ciência em cativeiro
8 – O nascimento do comércio
9 – O capitalismo do desastre

III

10 – Deuses e demônios
11 – Ápice
12 – Terra saqueada
13 – Os dois profetas
14 – Livramento

IV

15 – O gambito
16 – Mistério
17 – Acerto de contas

18 – A fórmula do império

EPÍLOGO: Castelo de cartas

POSFÁCIO

AGRADECIMENTOS
NOTAS
ÍNDICE
ÍNDICE DE NAVEGAÇÃO

# NOTA DA AUTORA

Este livro se baseia em mais de trezentas entrevistas com cerca de 260 pessoas, além de um vasto acervo de correspondências e documentos. A maioria foi conduzida especificamente para este livro. Algumas foram extraídas de meus últimos sete anos de reportagens sobre a OpenAI, a indústria da inteligência artificial e seus impactos globais para o *MIT Technology Review*, *The Wall Street Journal* e *The Atlantic*. Mais de 150 entrevistas foram realizadas com mais de noventa executivos e funcionários ou ex-funcionários da OpenAI, além de alguns que foram contratados e tiveram acesso à documentação detalhada sobre algumas práticas de desenvolvimento de modelos da empresa. Outra parte envolveu cerca de quarenta executivos e funcionários ou ex-funcionários das empresas Microsoft, Anthropic, Meta, Google, DeepMind e Scale, bem como pessoas próximas a Sam Altman.

Qualquer e-mail, documento ou mensagem do Slack citado provém de cópias ou capturas de tela de tais documentos e correspondências, ou aparece exatamente como consta em processos judiciais. Nos casos em que não possuo uma cópia, parafraseio o texto sem aspas. Há uma única exceção, indicada nas notas finais. Todos os diálogos foram reconstruídos a partir das memórias dos envolvidos e de anotações realizadas na época ou — quando indicado nas notas finais — extraídos de gravações de áudio ou transcrições. Na maioria dos casos, eu ou minha equipe de verificação de fatos pedimos às pessoas que relembravam as falas que repetissem ou confirmassem as informações novamente, com vários meses de intervalo, com o intuito de testar sua consistência. Cada momento, número, nome e codinome, bem como cada detalhe técnico sobre os modelos da OpenAI — como a composição dos dados de treinamento ou a quantidade de chips utilizados —, foi corroborado por ao menos duas pessoas, com anotações e documentação contemporâneas ou, em alguns casos específicos (também indicados nas notas finais), com reportagens de outros meios de comunicação. Isso vale para praticamente todos os outros detalhes sobre a OpenAI que constam neste livro. Se mencionei alguém pelo nome, não significa necessariamente que eu tenha conversado diretamente com essa pessoa. Quando me refiro a pensamentos ou sentimentos de alguém, é porque o indivíduo descreveu tal pensamento ou sentimento — a mim, a alguém com quem conversei, em um e-mail ou

gravação que obtive, ou em uma entrevista pública.

Este livro não é um título corporativo. Embora relate a história interna da OpenAI, esta narrativa serve como um prisma para enxergar muito além desta única empresa. É um perfil de uma ambição científica transformada em uma busca ideológica e agressiva, movida pelo dinheiro; um exame de sua pegada multifacetada e expansiva; uma meditação sobre o poder. Para tal propósito, durante minhas reportagens, passei um tempo considerável inserida em comunidades ao redor do mundo, ansiando por compreender suas histórias, culturas, vidas e experiências ao lidar com os impactos tão viscerais da IA. Espero que as histórias dessas pessoas se destaquem nestas páginas tanto quanto as narrativas que se desenrolam entre as paredes de uma das organizações mais secretas do Vale do Silício.

Entrei em contato com todas as figuras e empresas-chave descritas neste livro para solicitar entrevistas e comentários. Sam Altman e a OpenAI optaram por não cooperar.

PRÓLOGO

# UMA CORRIDA PELO TRONO

Na sexta-feira, 17 de novembro de 2023, por volta do meio-dia, horário do Pacífico, Sam Altman, CEO da OpenAI, garoto-propaganda do Vale do Silício e avatar da revolução da IA generativa, entrou em uma reunião no Google Meet e se deparou com quatro dos cinco integrantes do conselho da empresa o encarando.

Na tela de seu vídeo, Ilya Sutskever, membro do conselho e cientista-chefe da OpenAI, foi direto ao ponto: Altman estava sendo demitido, e o anúncio seria feito em breve.

Altman estava no quarto de um hotel de luxo em Las Vegas para assistir à primeira corrida de Fórmula 1 da cidade em décadas, um evento repleto de celebridades como Rihanna e David Beckham. Aquela era uma breve pausa no rigoroso calendário de viagens que ele mantinha desde que a empresa lançara o ChatGPT, cerca de um ano antes. Por um instante, Altman ficou tão surpreso que não conseguiu dizer nada. Desviou o olhar enquanto tentava recuperar a compostura. À medida que a conversa continuava, ele tentou, de seu jeito característico, amenizar a situação.

"Como posso ajudar?",[1] perguntou.

O conselho lhe disse para apoiar Mira Murati, a diretora interina que havia sido escolhida, que até então exercia o cargo de diretora de tecnologia da empresa. Ainda confuso[2] e se perguntando se aquilo seria um pesadelo, Altman concordou.

Alguns minutos depois, Sutskever enviou outro link do Google Meet para Greg Brockman, presidente da OpenAI e aliado próximo de Altman, que havia sido o único membro do conselho ausente na reunião anterior. Sutskever informou a Brockman que ele não faria mais parte do conselho, mas seu cargo na empresa seria mantido.

O anúncio público foi feito[3] logo em seguida. "A saída do sr. Altman segue um processo de revisão deliberativo conduzido pelo conselho, que encontrou inconsistências na franqueza de comunicações com seus membros, prejudicando sua competência de exercer as responsabilidades que lhe cabem. Portanto, o conselho não tem mais confiança em sua capacidade de continuar

liderando a OpenAI."

À primeira vista, a OpenAI estava no auge de seu poder. Desde o lançamento do ChatGPT, em novembro de 2022, a empresa havia se tornado a história de sucesso mais fantástica do Vale do Silício. O ChatGPT era o aplicativo de consumo com o maior crescimento na história. A valorização da startup seguia uma trajetória meteórica que fazia investidores salivarem e profissionais top de linha disputarem vagas na empresa, vista como um foguete em ascensão. Poucas semanas antes, a OpenAI havia sido avaliada em até 90 bilhões de dólares, como parte de uma oferta secundária que estava prestes a ser concluída, permitindo que funcionários vendessem suas ações aos ávidos investidores. Alguns dias antes, a empresa havia realizado um evento extremamente aguardado e celebrado, no qual apresentou sua linha mais ambiciosa de produtos até então.

Aos olhos do público, Altman era o homem que havia tornado tudo isso possível. Passara a primavera e o verão viajando pelo mundo, alcançando um nível de celebridade que levava a mídia a compará-lo a Taylor Swift. Encantara praticamente todos com o porte discreto, as declarações ousadas e a aparente sinceridade.

Antes de Las Vegas, Altman havia mais uma vez visitado diversos países, participando de um painel na APEC CEO Summit, onde fez discursos e deslumbrou seus ouvintes.

"Por que você dedica sua vida a este trabalho?", perguntou Laurene Powell Jobs, fundadora e presidente do Emerson Collective e viúva de Steve Jobs.

"Acho que esta será a tecnologia mais transformadora e auspiciosa já inventada pelo homem", respondeu ele. "Quatro vezes na história da OpenAI, a mais recente sendo há apenas algumas semanas, tive a oportunidade de estar presente quando, de certo modo, afastamos o véu da ignorância e avançamos a fronteira das descobertas, e poder fazer isso é, tipo, a honra profissional de uma vida inteira."

Os funcionários, em choque,[4] souberam da demissão de Altman da mesma forma que todo mundo, com o link do anúncio público sendo enviado de um a outro celular dentro da empresa. O que mais surpreendeu foi o abismo entre a notícia e a reputação de Altman. A empresa já contava com cerca de oitocentas pessoas. Na época, os funcionários tinham menos oportunidades de se encontrarem e interagirem pessoalmente com o CEO. Contudo, seu comportamento carismático nos palcos globais não diferia muito do jeito como se portava em reuniões gerais, em eventos da empresa e no escritório, quando não estava viajando.

À medida que os rumores se espalhavam e os funcionários rolavam freneticamente por seus feeds no X, antes conhecido como Twitter, em busca de qualquer informação, alguém no escritório se apegou à explicação que parecia mais lógica e gritou: "Altman está concorrendo à presidência!" Por um

momento, a tensão se aliviou, até que perceberam que não era o caso, e a especulação foi retomada com nova intensidade e apreensão. Será que Altman havia cometido alguma ilegalidade? Os funcionários especularam que talvez tivesse a ver com sua irmã. Um mês antes, ela havia alegado em tweets que viralizaram que o irmão havia abusado dela. Consideraram que talvez não fosse algo ilegal, mas eticamente questionável, por ventura ligado a outros investimentos de Altman ou a sua captação de recursos com investidores sauditas para uma nova iniciativa de chips de IA.

Sutskever publicou um anúncio no Slack da OpenAI. Em duas horas, realizaria uma reunião online com todos os funcionários para responder perguntas.

"Foram as duas horas mais longas de toda a minha vida", lembrou um deles.

Sutskever, Murati e os demais executivos da OpenAI apareceram na tela, lado a lado, rígidos e sem ensaio, enquanto a reunião geral era transmitida para funcionários presenciais e remotos.

Sutskever estava sério. Entre os funcionários, era conhecido como um pensador profundo e místico, expressando-se frequentemente por meio de termos espirituais com uma sinceridade que podia ser encantadora para alguns e desconcertante para outros. Também era brincalhão e tinha um coração gentil. No escritório, usava camisas com estampas de animais e adorava pintá-los — um gato fofinho, alpacas bonitinhas, um dragão gracioso cuspindo fogo — ao lado de rostos abstratos e objetos do dia a dia. Algumas de suas pinturas amadoras decoravam o espaço, incluindo um trio de flores que se abriam no mesmo formato do logotipo da OpenAI, símbolo do que ele sempre incentivava os funcionários a construírem: "Uma pluralidade de IAs que amam a humanidade."

Naquele momento, Sutskever tentava transmitir uma sensação de segurança aos funcionários ansiosos, que enviavam perguntas e mais perguntas por um documento online. No entanto, ele era um mensageiro imperfeito; não era alguém que se destacava em transmitir suas mensagens com clareza.

"Houve algum incidente específico[5] que levou a isso?", leu Murati em voz alta, dando início à lista de perguntas dos funcionários.

"Muitas das perguntas no documento são sobre detalhes", respondeu Sutskever. "O quê, quando, como, quem. Eu gostaria de poder falar mais sobre isso tudo, mas não posso." Ele acrescentou que qualquer curioso deveria ler o comunicado à imprensa. "Na verdade, esclarece muita coisa. Talvez seja melhor ler o texto mais de uma vez."

Os funcionários ficaram desconcertados com a resposta. Haviam acabado de receber uma notícia catastrófica. Sem dúvida, como as pessoas mais diretamente afetadas pela situação, eles mereciam mais explicações do que o público geral.

Murati leu mais algumas perguntas. Como aquilo afetaria a relação com a

Microsoft? Principal investidora e licenciada exclusiva das tecnologias da OpenAI, a empresa era também a única fornecedora de infraestrutura de computação. Sem ela, todo o trabalho da startup — a realização de pesquisas, os treinamentos de modelos de IA, os lançamentos de produtos — simplesmente cessaria. Murati respondeu que não esperava nenhum impacto. Haviam acabado de conversar com Satya Nadella, CEO da Microsoft, e Kevin Scott, diretor de tecnologia.

"Eles estão inteiramente comprometidos com nosso trabalho", disse ela.

E quanto à oferta secundária da OpenAI? Funcionários com certo tempo de casa haviam recebido a opção de vender parte de suas ações, o que poderia equivaler a milhões de dólares. A oferta estava tão próxima que muitos tinham planos para adquirir imóveis, ou até mesmo já tinham adquirido.

"A oferta… nós, hmm, vamos ver", respondeu Brad Lightcap, diretor de operações, hesitante. "Estou em contato com os investidores que lideram essa iniciativa e alguns de nossos maiores nomes já presentes no quadro societário. Todos reafirmaram apoio incondicional à empresa."

Após diversas outras perguntas serem respondidas vagamente, um funcionário tentou mais uma vez descobrir o que Sam havia feito. Será que estava relacionado com seu papel na empresa ou envolvia sua vida pessoal? Mais uma vez, Sutskever direcionou todos ao comunicado oficial.

"Na realidade, a resposta para esta pergunta está no texto", disse ele.

Murati continuou lendo o documento.

"As perguntas sobre detalhes serão respondidas em algum momento ou nunca?"

Sutskever respondeu:

"É melhor manterem suas expectativas baixas."

À medida que a reunião geral continuava e as respostas de Sutskever pareciam cada vez mais sem sentido, o desconforto dos funcionários rapidamente se transformou em raiva.

"Quando um grupo de pessoas passa por uma experiência difícil, elas geralmente acabam se tornando mais unidas e próximas umas das outras. Essa experiência difícil vai nos unir ainda mais como equipe e, portanto, nos tornaremos mais produtivos."

"Como você concilia o desejo de crescer juntos em meio a uma crise com uma frustrante falta de transparência?", escreveu um funcionário. "Normalmente, a verdade é uma condição primordial para a reconciliação."

"Bem, eu compreendo", respondeu Sutskever. "A situação não é perfeita."

Murati tentou conter a tensão crescente:

"A missão é muito maior do que qualquer um de nós."

Lightcap reforçou sua mensagem: os parceiros, clientes e investidores da OpenAI haviam enfatizado que continuavam alinhados com a missão da empresa.

"Se é que algo mudou, acho que agora temos um dever ainda maior de nos

empenharmos por essa missão."

Sutskever tentou novamente transmitir tranquilidade:

"Temos todos os ingredientes, todos eles: a computação, a pesquisa e as descobertas, que são incríveis. Quando se sentirem inseguros, receosos, lembrem-se dessas coisas. Visualizem mentalmente o tamanho do *cluster* e imaginem todas aquelas GPUs trabalhando juntas."

Um funcionário enviou uma nova pergunta.

"Devemos nos preocupar com uma tomada hostil de controle, por meio de uma influência coercitiva, dos atuais integrantes do conselho?", leu Sutskever em voz alta. "Tomada hostil de controle?", repetiu Sutskever, com um tom diferente na voz. "O conselho sem fins lucrativos da OpenAI agiu inteiramente de acordo com seus objetivos. Não se trata de uma tomada hostil, de modo algum. Discordo dessa pergunta."

Naquela noite, diversos funcionários se reuniram na casa de um colega para uma festa que havia sido planejada antes da demissão de Altman. Havia também convidados de outras empresas de IA, incluindo o Google DeepMind e a Anthropic.

Pouco antes do evento,[6] um aviso foi enviado a todos os participantes: "Estamos adicionando uma segunda sala temática para esta noite: 'Sala livre de OpenAI', onde não haverá nenhuma conversa sobre a OpenAI." No fim, poucos passaram muito tempo na nova sala, já que a maioria queria falar sobre o assunto.

Naquela tarde, Brockman havia anunciado que estava pedindo demissão como forma de protesto. Satya Nadella, da Microsoft[7] — que ficara furioso por ter sido comunicado sobre a demissão de Altman apenas minutos antes de ser efetivada —, publicou um tweet cuidadosamente redigido: "Temos um acordo de longo prazo com a OpenAI, com acesso total a tudo de que precisamos para cumprir nossa agenda de inovação e um roteiro empolgante de produtos; seguimos comprometidos com nossa parceria, e com Mira e sua equipe."

Enquanto os rumores continuavam a se espalhar, chegou a notícia de que mais três pesquisadores seniores haviam deixado a empresa: Jakub Pachocki e Szymon Sidor, funcionários das primeiras gerações e dois dos que tinham mais tempo de casa na OpenAI, além de Aleksander Mądry, professor do MIT que estava de licença e havia se juntado recentemente à companhia. As saídas alarmaram ainda mais os funcionários — um sinal de sangramento na liderança e na base de talentos, capaz de assustar investidores e paralisar a oferta de ações em andamento ou, pior ainda, arruinar a empresa.

Na festa, o clima ficou cada vez mais sombrio e tenso. O cancelamento da oferta significaria perder uma recompensa financeira considerável por todo o trabalho árduo realizado, sem mencionar a possibilidade de dissolução da própria empresa, que desperdiçaria tanto esforço e potencial acumulado.

Naquela mesma noite, o conselho e a liderança remanescentes da OpenAI

realizavam uma série de reuniões cada vez mais hostis. Após a reunião geral, a aparência de unidade entre Sutskever e os demais executivos havia desmoronado. Muitos dos que haviam se sentado ao lado dele durante a transmissão ao vivo foram quase tão surpreendidos quanto o restante da equipe, tomando conhecimento da demissão de Altman apenas instantes antes do anúncio. Irritados com o desempenho desastroso de Sutskever,[8] exigiram uma reunião com o restante do conselho. Cerca de uma dúzia de executivos, incluindo Murati e Lightcap, se reuniu em uma sala de conferências no escritório.

Sutskever participou virtualmente, com os três diretores independentes: Adam D'Angelo, cofundador e CEO do site de perguntas e respostas Quora; Tasha McCauley, empreendedora e pesquisadora associada sênior em gestão no instituto de políticas RAND; e Helen Toner, pesquisadora australiana do centro de estudos CSET (Centro de Segurança e Tecnologia Emergente) da Georgetown University.

Sob uma avalanche de perguntas, os quatro membros do conselho evitaram a todo instante oferecer mais explicações, alegando manter a confidencialidade devido a responsabilidades legais. Diversos líderes começaram a demonstrar abertamente sua indignação.

"Vocês estão afirmando que Sam não é confiável" disse, furiosa, Anna Makanju, vice-presidente de assuntos globais, que frequentemente acompanhava Altman em suas viagens diplomáticas corporativas ao redor do mundo. "Mas essa não é nossa experiência com ele."

A liderança reunida pressionou o conselho a renunciar e ceder seus assentos a três funcionários, ameaçando pedir demissão coletiva caso a demanda não fosse atendida de imediato. Jason Kwon, diretor de estratégia e ex-conselheiro jurídico da empresa, elevou a aposta. Afirmou que, na realidade, seria *ilegal* o conselho se recusar a renunciar, porque, caso a empresa desmoronasse, isso configuraria uma violação dos deveres fiduciários dos conselheiros.

Os integrantes discordaram, declarando terem consultado cuidadosamente advogados antes de optar pela demissão de Altman e sustentaram que haviam agido de acordo com as responsabilidades que lhes cabiam. Argumentaram que a OpenAI não era uma empresa comum e, portanto, seu conselho era inusitado. Ela tinha uma estrutura única, elaborada pelo próprio Altman, que concedia ao conselho ampla autoridade para agir não em prol dos acionistas, mas da missão: garantir que a IAG, ou inteligência artificial geral, beneficiasse a humanidade.

Altman havia, de mais a mais, exaltado por anos o poder do conselho de demiti-lo como o principal mecanismo de governança da OpenAI. Helen Toner reforçou o argumento:

"Se esta ação destruir a empresa, na verdade, ela ainda pode estar de acordo com a missão."

A liderança transmitiu as palavras de Toner aos funcionários em tempo

real. Ela não se importava se a empresa estivesse com os dias contados. Talvez, concluíram muitos, essa fosse justamente a intenção. Ao ouvir isso, e diante da possibilidade de perderem todo o valor de suas ações, uma pessoa na festa começou a chorar.

Na manhã seguinte, sábado,[9] 18 de novembro, dezenas de pessoas, incluindo funcionários da OpenAI, reuniram-se na mansão de 27 milhões de dólares de Altman para aguardar notícias.

Os três pesquisadores seniores que haviam pedido demissão — Pachocki, Sidor e Mądry — encontraram-se com Altman e Brockman para discutir a possibilidade de recriar a empresa e dar continuidade ao trabalho. Para alguns, a notícia de tais conversas aumentou a ansiedade: uma nova concorrente da OpenAI poderia intensificar ainda mais a instabilidade. Para outros, ofereceu uma luz no fim do túnel: caso Altman de fato fundasse uma nova empresa, eles o seguiriam.

A liderança remanescente da OpenAI deu ao conselho um prazo de até as 17 horas, horário do Pacífico, para agir: reintegrar Altman e renunciar ou enfrentar uma retirada em massa de funcionários. Os membros do conselho negaram. Durante o fim de semana, telefonaram freneticamente, às vezes de madrugada, para qualquer pessoa em sua rede de contatos que atendesse. Diante da crescente indignação da equipe e dos investidores com a demissão de Altman, Murati já não estava disposta a prosseguir como CEO interina. O conselho precisava substituí-la por alguém capaz de recuperar a estabilidade ou encontrar novos membros que conseguissem enfrentar Altman, caso ele de fato retornasse.

Naquela noite, após o prazo expirar, Jason Kwon enviou um memorando aos funcionários, que dizia: "Ainda estamos trabalhando[10] por uma resolução e permanecemos otimistas. Por resolução, digo trazer de volta Sam, Greg, Jakub, Szymon e Aleksander."

Altman tuitou em seu característico estilo, em letras minúsculas: "eu amo tanto a equipe da openai."

Dezenas de outros funcionários começaram a retuitar a mensagem de Altman com um emoji de coração.

No domingo, Altman e Brockman voltaram ao escritório para negociar seu retorno. Ao longo do dia, cada vez mais funcionários se juntaram a eles, esperando ansiosamente. A essa altura, a maioria dos empregados, da liderança e do próprio conselho mal havia dormido nas últimas 36 horas. Tudo começava a virar um borrão. Altman publicou uma selfie, com os lábios comprimidos e o cenho franzido, segurando um crachá de visitante. "essa é a primeira e a última vez que uso um desses", escreveu como legenda. A liderança definiu mais um prazo até as 17 horas para que o conselho reintegrasse Altman e renunciasse.

Naquele momento, a pressão vinha de todos os lados. A Microsoft, outros

investidores da OpenAI e figuras de peso do Vale do Silício manifestavam-se publicamente a favor de Altman. Uma fonte chegou a revelar o plano de ação[11] para a imprensa: caso a decisão não fosse revertida, os funcionários deixariam a empresa em massa, a Microsoft interromperia o acesso à infraestrutura computacional e os investidores entrariam com ações judiciais. A combinação de tais forças impossibilitaria a OpenAI de existir sem Altman.

Ainda assim, havia resistência. Perto das 21 horas, já muito além do último prazo, Sutskever postou uma longa mensagem no Slack em nome do conselho. Altman não voltaria, e Emmett Shear, ex-CEO da Twitch, seria o novo diretor-executivo interino da OpenAI. Ele e Sutskever chegariam ao escritório dentro de cinco minutos para apresentar a nova visão da empresa.

"O conselho mantém firmemente sua decisão[12] como o único caminho para avançar e defender a missão da OpenAI", escreveu. "Em suma, o comportamento de Sam e sua falta de transparência nas interações com o conselho minaram sua capacidade de supervisionar a empresa da maneira que deveria."

O Slack explodiu imediatamente com dezenas de respostas furiosas por parte dos funcionários:

"Vocês e mais quem, porra?"

"Vocês estão delirando."

"Emmett vai ser o CEO de nada."

Cerca de duzentos funcionários saíram do escritório em protesto, com o intuito de boicotar a reunião. Murati apressou os executivos para saírem do prédio. Quando Shear chegou acompanhado de Sutskever, havia pouco mais de uma dúzia de pessoas na plateia.

Anna Brockman, esposa de Greg,[13] aproximou-se de Sutskever, que quatro anos antes havia sido o celebrante de seu casamento civil. Em lágrimas, ela o abraçou com força e implorou que reconsiderasse sua posição.

Muitos dos funcionários que deixaram o escritório se reuniram em casas de colegas para passar a noite; centenas se juntaram a um grupo no aplicativo Signal para acompanhar as atualizações. Mais tarde naquela noite, Satya Nadella anunciou que estava contratando Altman e Brockman para liderarem uma nova divisão de IA. A novidade se espalhou rapidamente: qualquer funcionário que quisesse seguir trabalhando com ele teria um emprego garantido na Microsoft.

A notícia fez o clima mudar de apreensão para rebeldia. Com a percepção de uma alternativa segura em mãos, os funcionários sentiram que tinham uma vantagem para enfrentar o conselho e Shear. Na casa de um deles, lotada com mais de cem colegas da OpenAI, executivos e pesquisadores seniores redigiram uma carta aberta para aumentar a pressão, reiterando com mais convicção ainda a ameaça da liderança: sem a reintegração de Altman e a renúncia do conselho, todos poderiam pedir demissão imediatamente e se juntar à Microsoft.

O grupo trabalhou para divulgar a carta ao máximo, compartilhando-a em diversos canais privados e telefonando para funcionários que não estavam presentes para assiná-la. À medida que o documento atingia um número crítico de adesões, muitos outros se apressaram para assinar, temendo levantar suspeitas caso não o fizessem. Em menos de 24 horas, já contava com mais de setecentas assinaturas, de um total aproximado de 770 empregados. Dezenas enviaram e-mails idênticos ao conselho, um após o outro. Em peso, foram ao X postar a mesma mensagem: "A OpenAI não é nada sem sua equipe."

Então, no meio da madrugada, os funcionários perceberam o nome de Sutskever aparecer na carta aberta.

Pouco tempo depois, ele se pronunciou publicamente: "Lamento profundamente por minha participação nas ações do conselho", tuitou nas primeiras horas da segunda-feira. "Nunca foi minha intenção prejudicar a OpenAI. Amo tudo o que construímos juntos e farei o possível para reconciliar a empresa."

Na terça-feira, 21 de novembro, da casa de Altman, a liderança da OpenAI fez uma chamada com o conselho. Cinco dias depois, todos já sentiam o impacto da privação de sono e da exaustão. O Dia de Ação de Graças se aproximava, e o desespero de ambos os lados finalmente havia aberto caminho para uma resolução.

Ao longo do dia, todos os envolvidos começaram a considerar seriamente diferentes configurações.

Altman e Brockman, que a princípio insistiam em retornar à OpenAI com assentos no conselho, acabaram cedendo e concordando em não os receber. Por sua vez, o conselho, ao perceber que não havia como preservar a empresa sem Altman, finalmente aceitou sua reintegração.

Naquela noite, mais tarde, as partes chegaram a um acordo sobre os três assentos independentes do conselho. D'Angelo permaneceria como membro; Toner e McCauley deixariam seus cargos; Bret Taylor, que tinha sido co-CEO da Salesforce e era ex-CTO do Facebook, e Larry Summers, ex-secretário do Tesouro dos EUA e ex-presidente de Harvard, assumiriam as vagas. Como parte do acordo, o novo conselho se comprometeu a adicionar novos integrantes no futuro, e Altman aceitou se submeter a uma investigação.

Todos concordaram que o mais importante para a empresa naquele momento era transmitir unidade, estabilidade e reconciliação. Dois dias depois, Altman faria um post no X que pareceria ensaiado: "acabei de passar algumas horas muito agradáveis com @adamdangelo. feliz Dia de Ação de Graças, das nossas famílias para as suas 🦃". Dez dias depois, Brockman publicaria uma foto com Sutskever, ambos abraçados e sorrindo abertamente. No escritório, o artista residente[14] da empresa conectaria o DALL-E, gerador de imagens, a uma impressora colorida para criar pequenos adesivos em formato de coração e com desenhos caleidoscópicos. Ao lado da impressora, haveria um enorme coração cor-de-rosa com a frase: "A OpenAI

não é nada sem sua equipe."

Em dezembro, Altman descreveria toda a experiência em um podcast com Trevor Noah como o segundo pior momento de sua vida, superado apenas pelo falecimento do pai. No mês seguinte, em janeiro, a oferta secundária seria concluída, avaliando a OpenAI em 86 bilhões de dólares.

Entretanto, tudo isso ainda estava por vir. Na noite de terça-feira, 21 de novembro, o momento era apenas de celebração. Com o anúncio do retorno de Altman e o novo acordo, os funcionários voltaram em massa ao escritório para se abraçar, chorar e tocar música no volume máximo. Em algum momento, alguém[15] ligou uma máquina de fumaça, que disparou o alarme de incêndio; no entanto, ninguém se importou e a festa continuou.

Brockman tirou uma selfie com o grupo, uma foto que capturou o delírio eufórico e quase insano de quem acabara de sobreviver a uma crise. Publicou a imagem no X com a legenda:[16] "estamos mais do que de volta."

A notícia da demissão de Altman chegou a mim no meio de uma entrevista para este livro. Eu havia silenciado o celular, felizmente alheia à semana caótica que estava prestes a começar. Vinte minutos depois, toquei na tela para ver as horas e me deparei com uma enxurrada de notificações perdidas. Assim começou uma sequência nebulosa de dias repletos de adrenalina enquanto eu me apressava para entender o que estava acontecendo.

Nas semanas seguintes, amigos, familiares e a mídia me perguntariam dezenas de vezes: o que tudo isso significava, afinal? Teria sido apenas uma distração em nome do entretenimento? Ou será que haveria consequências para o restante de nós? Eu já acompanhava a OpenAI por cinco anos. Em 2019, fui a primeira jornalista a conseguir extenso acesso à empresa e a escrever seu primeiro perfil público. Para mim, aqueles acontecimentos não eram simples manobras de poder do Vale do Silício. O drama trazia à tona uma das questões mais urgentes de nossa geração: como controlar a inteligência artificial?

A IA é uma das tecnologias mais decisivas desta era. Em pouco mais de uma década, reformulou a espinha dorsal da internet, tornando-se uma mediadora onipresente das atividades digitais. Em ainda menos tempo, está a caminho de reconfigurar uma série de outras funções vitais da sociedade, que vão desde a saúde à educação, das leis às finanças, do jornalismo ao governo. O futuro da IA — a forma que essa tecnologia assumirá — está intimamente ligado ao nosso próprio futuro. A questão de como controlá-la é, no fundo, a questão de como garantir que nosso futuro será melhor, não pior.

Desde o começo, a OpenAI se apresentou como um experimento audacioso em busca dessa resposta. Foi fundada por um grupo que incluía Elon Musk e Sam Altman, com o apoio de outros bilionários como Peter Thiel, para ser mais do que um laboratório de pesquisa ou uma empresa qualquer. Os fundadores afirmavam o compromisso radical de desenvolver a chamada inteligência artificial geral (IAG), a forma mais poderosa de IA já

concebida, não para o lucro dos acionistas, mas para o benefício da humanidade. Para isso, Musk e Altman estruturaram a empresa como uma organização sem fins lucrativos, comprometendo-se com um bilhão de dólares para seu funcionamento. A OpenAI não desenvolveria produtos comerciais; seria integralmente dedicada à pesquisa, movida apenas pelas intenções mais puras de instaurar uma IAG que revelaria uma utopia global, e não o oposto. Musk e Altman também prometeram compartilhar o máximo possível de suas pesquisas e colaborar amplamente com outras instituições. Se o objetivo era fazer o bem ao mundo, a abertura — por isso o nome *Open*AI — e a participação democrática no desenvolvimento tecnológico seriam essenciais. Alguns anos depois, a liderança acabou indo mais longe, fazendo uma promessa de autossacrifício, caso necessário: "Estamos preocupados que o desenvolvimento tardio da IAG possa se tornar uma corrida competitiva sem tempo para precauções adequadas em relação à segurança", escreveram. Se outra iniciativa de uma IAG benigna ultrapassar o progresso da OpenAI, "nós nos comprometemos a parar de competir e a começar a colaborar com tal projeto".

Contudo, quando comecei a perfilar a OpenAI, seu compromisso com tais ideais já estava se dissolvendo rapidamente. Em apenas um ano e meio desde sua fundação, os executivos perceberam que o caminho que desejavam seguir no desenvolvimento da IA exigiria valores exorbitantes de dinheiro. Musk e Altman, que até então haviam atuado de forma relativamente distante como copresidentes, tentaram se instalar como CEO. Altman venceu, e Musk deixou a organização no início de 2018, levando consigo seu investimento. Em retrospecto, a ruptura foi o primeiro grande sinal de que a OpenAI não era, de fato, um projeto altruísta, mas sim movido por ego.

A perda de seu principal financiador lançou a empresa em um caminho financeiramente incerto. Para remediar a situação, Altman reformulou a estrutura jurídica da OpenAI. No interior da entidade sem fins lucrativos, criou um braço com fins lucrativos, a OpenAI LP, a fim de levantar capital, comercializar produtos e oferecer retorno a investidores, como qualquer outra empresa. Quatro meses depois, em julho de 2019, a OpenAI anunciou um novo financiador de um bilhão de dólares: Microsoft, a gigante de software e serviços de nuvem.

Pouco tempo depois, em agosto de 2019, cheguei aos escritórios da OpenAI pela primeira vez. Após três dias absorta entre funcionários e dezenas de entrevistas, pude perceber que o experimento de governança idealista estava desmoronando. A empresa havia se tornado competitiva, tomada de segredos e isolamento, até mesmo receosa a respeito do mundo exterior, intoxicada pelo poder de controlar uma tecnologia tão fenomenal. As noções de transparência e democracia, de autossacrifício e colaboração, haviam desaparecido. Seus executivos tinham uma única obsessão: serem os primeiros a alcançar a inteligência artificial geral, e fazê-la à sua própria imagem.

Nos quatro anos seguintes, a OpenAI se tornou tudo aquilo que dizia que

jamais seria. Virou uma organização sem fins lucrativos apenas no nome, comercializando agressivamente produtos como o ChatGPT e buscando avaliações de mercado nunca vistas. Tornou-se ainda mais cheia de segredos, não apenas restringindo o acesso às próprias pesquisas, mas também alterando as normas de toda a indústria para que uma parcela significativa do desenvolvimento de IA passasse a acontecer longe do escrutínio público. Sendo assim, desencadeou exatamente a corrida destrutiva sobre a qual alertou, acelerando brutalmente a comercialização e o uso da tecnologia sem antes corrigir suas falhas tão perigosas ou as formas como poderia amplificar e explorar as fraturas de nossa sociedade. Nesse percurso, os conflitos entre líderes e funcionários se intensificaram, ao passo que diferentes grupos dentro da empresa tentavam tomar o controle e reformar a OpenAI de acordo com as próprias visões.

A queda e o retorno de Sam Altman, em novembro de 2023, foram a prova definitiva de que o experimento de controle havia fracassado. Não apenas porque o conselho sem fins lucrativos da OpenAI cedeu às pressões dos interesses financeiros, dissolvendo o último resquício da fachada altruísta da empresa, mas porque ilustrou claramente o quanto o futuro da IA está sendo moldado por uma luta de poder entre uma parcela de elites do Vale do Silício. Mesmo que os acontecimentos tivessem tomado outro caminho e o conselho tivesse conseguido substituir Altman, nada alteraria o fato de que uma decisão tão importante foi tomada a portas fechadas. Fora desse círculo restrito de tecnófilos bilionários, seus rivais ideológicos mais ferrenhos e um gigante tecnológico multibilionário, e até mesmo os próprios funcionários da OpenAI, se viram, em sua maioria, no escuro quanto ao destino que os esperava.

Comecei a cobrir o campo de inteligência artificial muito antes de a OpenAI e o ChatGPT se tornarem sinônimos dessa tecnologia. Acompanhei sua evolução através do processo caótico da ciência e da inovação, enquanto pesquisadores testavam novas ideias, apresentavam seus maiores sucessos em conferências lotadas e os transformavam em produtos comerciais nas maiores empresas do mundo, como Google, Facebook, Alibaba e Baidu. Li centenas de artigos técnicos e entrevistei cientistas, engenheiros e executivos a fim de compreender suas visões de mundo e decisões — e como essas escolhas deixaram marcas digitais na forma como a tecnologia foi projetada e aplicada.* Conforme a marca digital da IA se expandia pelo mundo, observei as formas sutis e dramáticas pelas quais transformava vidas e comunidades. Viajei por cinco continentes para ouvir pessoas sobre essas experiências. Na Colômbia e no Quênia, conheci indivíduos que, diante de crises econômicas, recorreram à anotação de dados para a indústria da IA, somente para acabarem descobrindo que trabalhavam em condições semelhantes à da servidão por dívida. No Arizona e no Chile, encontrei-me com políticos locais e ativistas preocupados com o crescimento sombrio das metrópoles de data centers, que sugavam os preciosos recursos hídricos de suas cidades.

Por meio de minhas reportagens, cheguei a duas conclusões. A primeira

foi que a inteligência artificial é uma tecnologia que assume diversas formas. Na realidade, é um aglomerado de tecnologias que se transformam e evoluem não somente com base em méritos técnicos, mas também com o impulso das ideologias de seus criadores, e com os ventos da promoção e da comercialização. Embora o ChatGPT e outros supostos modelos de linguagem de grande escala ou aplicativos de IA generativa estejam agora sob a luz dos holofotes, são apenas uma manifestação da IA, que representa uma visão particular e incrivelmente estreita sobre como o mundo é e deveria ser. Nada sobre seu surgimento ou sua existência foi inevitável, mas sim o resultado de milhares de escolhas subjetivas, feitas por pessoas que tinham o poder de decisão. Do mesmo modo, as gerações futuras das tecnologias de IA não estão predeterminadas, mas volta a questão do controle: quem terá o poder de moldá-las?

A outra lição que aprendi foi que a manifestação atual da IA, bem como a trajetória de seu desenvolvimento, está seguindo uma direção alarmante. Superficialmente, a IA generativa é empolgante: uma ajuda criativa para gerar ideias e textos instantâneos; uma companhia para conversar tarde da noite e afastar a solidão; uma ferramenta que talvez, um dia, seja tão eficaz em impulsionar a produtividade que aumentará o faturamento total. Todavia, assim como antes pensávamos no Facebook somente como um espaço para publicar fotos de férias, reencontrar amigos de infância ou fomentar movimentos sociais positivos e transformadores, há muito mais por trás dessa fachada elegante e hipnotizante do que se constata à primeira vista. Por baixo de tudo isso, os modelos de IA generativa são monstruosidades construídas a partir do consumo de quantidades de dados, trabalho, poder computacional e recursos naturais anteriormente inconcebíveis. O GPT-4, sucessor do primeiro ChatGPT, é, segundo determinada medida, mais de quinze mil vezes maior do que o GPT-1, a primeira geração, lançada cinco anos antes. Os custos humanos e materiais exorbitantes recaem sobre amplas camadas da sociedade, principalmente as mais vulneráveis — pessoas que encontrei ao redor do mundo, trabalhadores e moradores de áreas rurais ou comunidades carentes da divisão norte-sul —, todas sofrendo com novos níveis de precariedade. Raramente elas sentiram qualquer benefício dessa chamada revolução tecnológica, já que os ganhos da IA generativa se concentram principalmente nas camadas mais altas.

Ao longo dos anos, encontrei apenas uma metáfora capaz de encapsular a natureza desses atores tão poderosos da IA: impérios. Durante a longa era do colonialismo europeu, os impérios tomavam e extraíam recursos que não lhes pertenciam e exploravam o trabalho das pessoas que subjugavam para minerar, cultivar e refinar tais recursos em prol do próprio enriquecimento. Projetavam ideias racistas e desumanizadoras sobre sua superioridade e modernidade a fim de justificar — e até mesmo atrair a aceitação dos conquistados — a invasão da soberania, o roubo e a submissão. Justificavam sua busca por poder pela necessidade de competir com outros impérios em uma corrida armamentista,

na qual todas as apostas estavam abertas. Em última análise, tudo isso servia para consolidar o poder de cada império e impulsionar sua expansão e seu progresso. Em termos simples, acumulavam riquezas extraordinárias ao longo do espaço e do tempo, impondo uma ordem colonial mundial, a grandes custos para todos os demais.

Os impérios da inteligência artificial não estão envolvidos na mesma violência e brutalidade explícitas daqueles que marcaram a história. No entanto, também se apoderam e extraem recursos preciosos para alimentar sua visão: o trabalho de artistas e escritores; os dados de incontáveis indivíduos que publicam suas experiências e observações na internet; a terra, a energia e a água necessárias para abrigar e operar imensos data centers e supercomputadores. Do mesmo modo, os novos impérios exploram o trabalho de pessoas ao redor do mundo para limpar, organizar e preparar tais dados, que se tornarão tecnologias lucrativas de IA. Projetam ideias sedutoras de modernidade e adotam uma postura agressiva quanto à necessidade de derrotar outros impérios, a fim de justificar e fomentar invasões de privacidade, roubos e a automação catastrófica de amplas e significativas áreas de oportunidades econômicas.

Atualmente, a OpenAI lidera nossa aceleração rumo a essa nova ordem mundial colonial dos tempos modernos. Em busca de uma visão amorfa de progresso, sua pressão agressiva sobre os limites de escala estabeleceu regras para uma nova era de desenvolvimento da inteligência artificial. Agora, cada gigante da tecnologia corre para superar os demais em escala, gastando somas tão astronômicas que até mesmo fazem com que se vejam obrigados a redistribuir e consolidar os próprios recursos. Na época em que a Microsoft investiu 10 bilhões de dólares na OpenAI, a empresa demitiu dez mil funcionários para reduzir os custos. Depois que o Google viu a OpenAI ultrapassá-lo, centralizou seus laboratórios de IA sob o nome Google DeepMind. À medida que a Baidu se apressava para desenvolver[17] seu equivalente ao ChatGPT, funcionários que trabalhavam com tecnologias de IA em prol do avanço na descoberta de novos medicamentos tiveram que suspender suas pesquisas e ceder os chips de seus computadores para desenvolver o chatbot. O paradigma atual da IA também está sufocando caminhos alternativos de desenvolvimento. O número de pesquisadores independentes não afiliados nem financiados pela indústria tecnológica diminuiu rapidamente, reduzindo a diversidade de ideias no campo que não estejam vinculadas com o lucro comercial de curto prazo. As próprias empresas, que antes investiam na ampliação de pesquisas exploratórias, já não podem mais arcar com isso diante dos custos exorbitantes do desenvolvimento da IA generativa. As novas gerações de cientistas estão se conformando com o novo *status quo* para se tornarem mais empregáveis. O que antes era sem precedentes tornou-se a norma.

Os impérios nunca foram tão ricos como agora. Quando terminei de escrever este livro, em janeiro de 2025, a OpenAI havia alcançado uma

avaliação de 157 bilhões de dólares. A Anthropic, uma concorrente, estava prestes a fechar um acordo que a avaliaria em 60 bilhões. Depois de firmar parceria com a OpenAI, a Microsoft triplicou seu capital de mercado, ultrapassando 3 trilhões de dólares. Desde o lançamento do ChatGPT, as seis[18] maiores gigantes da tecnologia viram, somadas, suas valorizações de mercado aumentarem em 8 trilhões de dólares. Ao mesmo tempo, cada vez mais dúvidas crescem sobre o real valor econômico da IA generativa. Em junho de 2024, um relatório do Goldman Sachs[19] reportou que os gastos com o desenvolvimento dessa tecnologia deveriam atingir 1 trilhão de dólares em poucos anos, até agora com "muito pouco a mostrar" em retorno. No mês seguinte, uma pesquisa[20] do Upwork Research Institute, com 2.500 trabalhadores em todo o mundo, revelou que, embora 96% dos executivos de alto escalão esperassem que a IA generativa aumentasse a produtividade, 77% dos funcionários que de fato utilizavam as ferramentas relatavam que elas, na realidade, aumentaram sua carga de trabalho, em parte pelo tempo gasto revisando o conteúdo gerado pela IA, em parte devido ao crescimento das demandas de seus superiores por mais trabalho. Em um artigo publicado pela *Bloomberg* em novembro, analisando o impacto financeiro da IA generativa, as jornalistas Parmy Olson e Carolyn Silverman resumiram com precisão: os dados "levantam uma perspectiva desconfortável[21] — a de que essa tecnologia supostamente revolucionária talvez nunca cumpra sua promessa de uma transformação econômica ampla, mas, em vez disso, somente concentre ainda mais riqueza no topo da pirâmide".

Enquanto isso, o restante do mundo está começando a desmoronar sob o peso dos custos humanos e materiais desta nova era. Trabalhadores no Quênia receberam salários miseráveis para filtrar violência e discurso de ódio das tecnologias da OpenAI, incluindo o ChatGPT. Artistas estão sendo substituídos pelos próprios modelos de IA que foram construídos a partir de suas obras, sem consentimento ou compensação. A indústria jornalística está atrofiando, ao passo que as tecnologias de IA generativa produzem volumes cada vez maiores de desinformação. Diante de nossos olhos, estamos assistindo à repetição de uma história antiga — e isso é apenas o começo.

A OpenAI não está desacelerando. Continua a perseguir escalas ainda maiores, com recursos sem precedentes, e o restante da indústria segue o mesmo caminho. Para conter as crescentes preocupações sobre o desempenho atual da IA generativa, Altman tem proclamado ainda mais alto os benefícios futuros da IAG. Em uma publicação no blog, em setembro de 2024,[22] declarou que a "Era da Inteligência", caracterizada por uma "prosperidade abundante", logo estaria diante de nós, com a superinteligência talvez surgindo em "alguns milhares de dias". "Acredito que o futuro será tão brilhante que ninguém conseguirá lhe fazer justiça tentando descrevê-lo agora", escreveu. "Embora ocorram de maneira gradual, triunfos extraordinários, como a resolução da crise climática, o estabelecimento de uma colônia espacial e a descoberta de toda a física, por fim se tornarão comuns." Neste ponto, a IAG

é em grande parte retórica — uma fantasia, um pretexto multifuncional para a OpenAI continuar a buscar cada vez mais riqueza e poder. Poucos outros possuem capital comparável para investir em alternativas. A empresa e seu pequeno grupo de concorrentes terão um oligopólio sobre a tecnologia, o qual estão nos vendendo como a chave para o futuro; qualquer um, seja empresa ou governo, que queira uma fatia dessa visão terá que depender desses impérios para obtê-la.

Há um caminho alternativo possível. A inteligência artificial não precisa ser o que é atualmente. Não precisamos aceitar a lógica de escala e consumo inéditos para alcançar avanço e progresso. Grande parte do que nossa sociedade realmente precisa — saúde e educação melhores, ar e água limpos e uma transição mais rápida para o fim dos combustíveis fósseis — pode ser assistida e impulsionada por modelos de IA significativamente menores e uma gama de outras abordagens. A própria IA também não será suficiente. Precisaremos de mais coesão social e cooperação global, justamente algumas das coisas que estão sendo desafiadas pela visão atual de desenvolvimento da IA.

Entretanto, os impérios da inteligência artificial não abrirão mão do poder tão facilmente. O restante de nós precisará retomar o controle sobre o futuro dessa tecnologia, e estamos em um momento crucial, no qual isso ainda é possível. Assim como os impérios do passado acabaram sucumbindo diante de formas mais inclusivas de governança, nós também podemos moldar juntos o futuro da IA. Os formuladores de políticas públicas podem implementar regras rigorosas de privacidade e transparência de dados, além de atualizar as proteções de propriedade intelectual para devolver às pessoas o controle sobre seus dados e trabalho. As organizações de direitos humanos podem promover normas e leis trabalhistas internacionais que garantam aos rotuladores de dados salários mínimos e condições de trabalho humanas, bem como reforçar os direitos trabalhistas e proporcionar o acesso a oportunidades econômicas dignas em todos os setores e indústrias. As agências de fomento podem estimular uma diversidade renovada na pesquisa em IA, a fim de desenvolver novas e fundamentais manifestações do que essa tecnologia pode ser. Por fim, todos nós podemos resistir às narrativas que a OpenAI e a indústria da inteligência artificial nos contaram para esconder os crescentes custos sociais e ambientais da tecnologia atrás de uma visão ilusória de progresso.

---

* Uma nota sobre as convenções de artigos científicos e o processo de revisão por pares: no campo da IA, os pesquisadores costumam publicar seus trabalhos diretamente online, em um repositório livre e aberto chamado arXiv (pronuncia-se *archive*), e eles podem ou não submetê-los a um processo de revisão por pares em uma conferência ou publicação, muitas vezes meses, ou até anos, depois, ou simplesmente optam por não o fazer. Tal prática se tornou tão corriqueira que muitos na área citam trabalhos com base em seu

impacto, e não por terem passado ou não pela revisão por pares. Neste livro, seguirei o mesmo padrão. As notas finais indicarão quais artigos não foram revisados, permanecendo apenas como pré-publicações.

I

CAPÍTULO 1

# DIREITO DIVINO

Todos os outros já haviam chegado,[1] mas Elon Musk, como sempre, estava atrasado.

Era o verão[2] de 2015, e um grupo de homens tinha se reunido para um jantar privado a convite de Sam Altman a fim de discutir o futuro da inteligência artificial e da humanidade.

Musk conhecera Altman, catorze anos mais novo, pouco antes e tivera uma boa impressão. Presidente da Y Combinator, renomada aceleradora de startups do Vale do Silício, Altman já tinha uma reputação consolidada. Após fundar sua primeira empresa aos dezenove anos, ele rapidamente se firmara no Vale como um estrategista e negociador brilhante, com ambições grandiosas, mesmo para um local tão conhecido por seus fundadores e suas ousadas ideias. Musk o considerava inteligente, determinado e, o mais importante, alguém que compartilhava de suas visões sobre a necessidade de desenvolver e controlar a inteligência artificial com cautela. Era como se, diria Musk[3] anos mais tarde em um processo judicial, Altman tivesse espelhado tudo o que ele já dissera sobre o assunto para conquistar sua confiança.

Da parte de Altman,[4] ele costumava dizer que Musk fora como um herói de infância. Após ter sido levado para conhecer a imensa fábrica da SpaceX em Hawthorne, na Califórnia, sua admiração apenas se intensificou. "O que ficou na memória[5] foi a expressão de absoluta certeza em seu rosto quando falava sobre enviar grandes foguetes a Marte", escreveu Altman mais tarde sobre a experiência. "Saí de lá pensando: 'Ah, então é isso que as pessoas chamam de convicção.'"

Havia algum tempo que Musk já vinha se sentindo profundamente preocupado com a inteligência artificial. Em 2012, conheceu Demis Hassabis, o professoral CEO do laboratório londrino DeepMind Technologies. Pouco depois, Hassabis também o visitou na fábrica da SpaceX. Enquanto ambos conversavam no refeitório, cercados pelo som de enormes peças de foguetes sendo transportadas e montadas, Hassabis levantou a hipótese de que uma IA mais avançada, do tipo que um dia seria capaz de superar a inteligência humana, poderia representar uma ameaça à humanidade. Ademais, o plano de

segurança de Musk, de colonizar Marte como forma de fuga, não funcionaria nesse cenário. A superinteligência, disse Hassabis com certo humor, simplesmente seguiria os humanos pela galáxia. Musk, sem achar graça alguma, investiu 5 milhões de dólares na DeepMind para manter a empresa sob vigilância.

Mais adiante, em sua festa de aniversário em 2013, nos exuberantes vinhedos do Vale do Napa, Musk se envolveu em um debate acalorado e sentimental com Larry Page, um velho amigo e cofundador do Google, sobre a possibilidade de a inteligência artificial superar a humana ser realmente um problema. Page não achava que fosse o caso e chamava isso de o próximo estágio da evolução. Quando Musk reagiu com relutância, Page o acusou de ser "especista" por discriminar espécies não humanas.

Depois disso, Musk começou a falar sem parar sobre o risco existencial da IA. Em um simpósio do MIT, descreveu a inteligência artificial como provavelmente a "maior ameaça existencial" à humanidade e seu desenvolvimento como "a invocação de um demônio". Encontrou-se com editores em Nova York, dominado pela ideia de escrever o próprio livro sobre ameaças de extinção, incluindo a IA. Mais tarde, em um evento periódico[6] chamado AI Salon, em Stanford, uma jovem pesquisadora chamada Timnit Gebru se aproximou dele após uma palestra e perguntou por que era tão obcecado pela IA, quando a ameaça das mudanças climáticas parecia ser mais claramente existencial. "As mudanças climáticas são ruins, mas não matarão todo mundo", respondeu Musk. "A IA tem o poder de extinguir a humanidade."

No final de 2013, quando soube que o Google compraria a DeepMind, Musk convenceu-se de que a união terminaria muito mal. Publicamente, alertou que, se o Google desse a uma hipotética IAG o objetivo de maximizar lucros, o software poderia buscar eliminar os concorrentes de uma empresa a qualquer custo. "Aniquilar todos os pesquisadores[7] de IA concorrentes como primeiro passo me parece uma ligeira falha de caráter", disse a *New Yorker*. Em uma chamada via Skype, com duração de mais de uma hora e realizada em um armário no andar de cima durante uma festa em Los Angeles, ele incentivou Hassabis a reconsiderar o acordo. "O futuro da IA",[8] disse Musk, "não deve ser controlado por Larry." No entanto, embora Musk não estivesse ciente, o Google já havia enviado uma equipe de pesquisadores de inteligência artificial em um jato particular aos escritórios da DeepMind para avaliar a aquisição. Como parte da aferição,[9] Jeff Dean, um dos primeiros e mais experientes funcionários do Google, revisou pessoalmente uma amostra do código da empresa e aprovou o acordo. Em janeiro de 2014, o Google confirmou a aquisição, que teria ocorrido por um valor entre 400 e 650 milhões de dólares.

Musk começou a organizar os próprios jantares com o intuito de discutir meios para enfrentar o Google. No início de 2015, também se reuniu com Barack Obama, então presidente dos EUA, para explicar os perigos da inteligência artificial e como torná-la mais segura e regulá-la. Na mesma época

voltou a encontrar Hassabis na SpaceX, desta vez para a primeira reunião do Conselho de Ética em IA da Google DeepMind, uma estrutura de governança que Page e Hassabis haviam proposto para supervisionar o desenvolvimento responsável das tecnologias da empresa. A reunião convenceu Musk[10] de que o conselho era uma farsa e inflamou suas preocupações, transformando-as em uma obsessão total em busca de contrariar a visão de Hassabis.

Nos anos seguintes, Musk passou a frequentemente descrever o executivo como um supervilão que precisava ser detido. Deixou inequivocamente claro que a OpenAI representava o bem diante do mal da DeepMind. No verão de 2016, pouco após a fundação da OpenAI, diversos funcionários conheceram Hassabis e relataram ao escritório: a DeepMind de fato pretendia dominar o mundo, e a descrição de Musk parecia correta. Um ano depois, Musk organizou uma reunião externa para funcionários da OpenAI na fábrica da SpaceX e começou um desabafo sobre Hassabis. Antes de fundar a DeepMind, ele havia passado sete anos administrando um estúdio de design de videogames de criação própria. "Hassabis literalmente desenvolveu um jogo em que um gênio maligno tenta criar uma IA para dominar o mundo", gritou Musk, referindo-se ao *Evil Genius*, título de 2004, "e as pessoas não enxergam isso. Não enxergam! E Larry? Larry acha que controla Demis, mas está ocupado demais fazendo windsurfe para perceber que ele está reunindo todo o poder."

A paranoia de Musk em relação a Hassabis se tornou fonte de entretenimento para os funcionários da DeepMind. Hassabis era incrivelmente ambicioso e podia ser intenso, é claro, mas também era gentil e comedido. "A criação da OpenAI parecia uma reação meio histérica a um homem de temperamento relativamente calmo", lembra um ex-pesquisador da DeepMind. "Parecia um pouco absurdo."[11]

Na lista de livros recomendados por Musk, estava *Superinteligência: caminhos, perigos, estratégias para um novo mundo*, no qual Nick Bostrom, filósofo de Oxford, argumenta que, se a inteligência artificial algum dia se tornasse mais sábia do que os humanos, seria algo difícil de controlar e poderia causar uma catástrofe existencial. Diante de um objetivo simples,[12] como produzir clipes de papel, essa IA superior poderia determinar que os homens representam uma ameaça à sua meta, pois demandam recursos necessários para tal. Bostrom então propôs uma solução: seria possível evitar o problema do controle da superinteligência "alinhando" a IA aos valores humanos — dando-lhe a capacidade de ir além das instruções explícitas para alcançar seus objetivos sem prejudicar pessoas. Tal ideia constituiu a base da disciplina de pesquisa na padronização de inteligência artificial que a OpenAI viria a defender. Para seus seguidores no Twitter,[13] Musk afirmou que "vale a pena ler" o livro.

Em janeiro de 2023, o ressurgimento de um e-mail que Bostrom havia

enviado a um LISTSERV em meados da década de 1990 fez as pessoas questionarem os valores humanos dele. “Sempre gostei da forma de pensar e falar implacavelmente objetiva”, escreveu. “Como exemplo, pensemos na seguinte frase: negros são mais estúpidos do que brancos. Gosto dela e acho que é verdadeira.” Bostrom se desculparia,[14] chamando o e-mail de “asqueroso” e uma representação imprecisa de suas opiniões.

Para Musk, Altman parecia um companheiro de jornada, alguém que carregava uma predisposição própria a se preparar para catástrofes. Em 2016, Altman contou a Tad Friend, veterano cronista Califórniano, da *New Yorker*, que, no caso de um cenário sobrevivencialista, planejava escapar para a Nova Zelândia com o investidor bilionário Peter Thiel, seu amigo próximo e mentor. Thiel descreveu Altman no mesmo artigo como “culturalmente muito judeu — um otimista, mas também um sobrevivencialista, cuja percepção é de que as coisas sempre podem dar muito errado”. Dois anos depois, Altman diria[15] à *Bloomberg* que estava brincando, mas que ainda mantinha uma mochila de emergência pronta. Estava particularmente preocupado com vírus biológicos novos e havia incluído máscaras de gás, antibióticos, água, baterias, uma barraca e uma arma. Contudo, em seu blog, em fevereiro de 2015, concordou com Musk que a superinteligência era “supostamente a maior ameaça[16] à existência e continuidade da humanidade”. Embora um vírus artificial devastador fosse mais provável de acontecer, ele afirmou que era “improvável que destruísse todos os humanos do universo”. “A propósito”, escreveu entre parênteses, “o *Superinteligência*, excelente livro de Nick Bostrom, é a melhor coisa que já li sobre o tema. Vale muito a pena ler”.

Alguns meses depois, em maio de 2015, Altman enviou um e-mail para Musk. “Tenho pensado muito[17] se é possível impedir que a humanidade desenvolva a IA”, escreveu. “Acho que a resposta é quase certamente não. Se isso vai acontecer de qualquer maneira, seria bom que alguém que não fosse o Google o faça primeiro.” Ele propôs que a Y Combinator, ou YC, como era conhecida, iniciasse um “Projeto Manhattan para IA”, estruturado “de forma que a tecnologia pertença ao mundo por meio de algum tipo de organização sem fins lucrativos”. “Obviamente, cumpriríamos/apoiaríamos agressivamente toda a regulamentação”, acrescentou, referindo-se às iniciativas recentes de Musk por supervisão governamental.

Musk respondeu que “provavelmente vale uma conversa”.

Em junho, Altman enviou outro e-mail com mais detalhes. “A missão seria criar a primeira IA geral e utilizá-la para o empoderamento individual — ou seja, a versão compartilhada do futuro que parece mais segura. Por norma, a segurança deve ser um requisito primordial.” Então propôs uma estrutura de governança que colocaria a ele e a Musk em posição de decisão. Ambos teriam assentos no conselho e convidariam mais três pessoas para se juntar a eles. “A tecnologia seria propriedade da organização e utilizada ‘para o bem do mundo’, e nos casos em que não ficasse claro como isso deveria ser aplicado, nós cinco decidiríamos”, disse Altman.

Caso Musk também pudesse se comprometer a se reunir com a equipe mensalmente, continuou Altman, isso ajudaria a "atrair as melhores pessoas para integrar o projeto". Se não tivesse tempo, seu endosso público "provavelmente ainda seria muito útil para o recrutamento".

Musk respondeu: "Concordo com tudo."

Altman então convidou Musk para o jantar privado sobre o futuro da inteligência artificial e da humanidade, com o intuito de que ele conhecesse um grupo de engenheiros e pesquisadores de IA de alto nível, que Altman esperava recrutar para o projeto. Com a confirmação de presença de Musk, o local do jantar passou a ser um restaurante em um dos pontos preferidos do fundador da SpaceX: o sofisticado Rosewood Hotel, de aproximadamente 65 mil metros quadrados e o valor de mil dólares por noite, cercado por dezenas de empresas de capital de risco ao longo da pitoresca e arborizada Sand Hill Road, que atravessa o Vale do Silício. A sala de jantar privativa onde se reuniram tinha uma varanda com vista para uma bela piscina cercada por ciprestes italianos e rosas de jardim. Quando Musk entrou, com mais de uma hora de atraso, os outros homens aguardavam ansiosamente. Entre eles estavam Altman, Greg Brockman, Dario Amodei e Ilya Sutskever.

Os integrantes do grupo logo se tornariam os principais líderes da organização sem fins lucrativos. Para capturar o espírito de sua missão em comum, Musk a batizaria de OpenAI. Com o tempo, quase todos os homens acabariam deixando a organização após conflitos com Altman e sua visão sobre inteligência artificial.

Quando Altman e Musk cortaram relações, e Altman o substituiu como o novo "queridinho" do Vale do Silício, ele mudou publicamente o discurso sobre suas crenças a respeito dos perigos do que estava construindo. "Agora estou definitivamente[18] do lado de quem diz que a IA será uma ferramenta", disse à *Business Insider* em 2023, "embora ache que os humanos do futuro e a sociedade serão extremamente diferentes, e temos a chance de sermos cuidadosos sobre como projetar esse futuro."

Musk passou a sentir que Altman o havia usado para catapultar a própria notoriedade.

Esse foi o eco de uma observação que acompanhou Altman ao longo da vida. "Você poderia lançá-lo de paraquedas[19] em uma ilha cheia de canibais e, cinco anos depois, ele seria o rei", disse certa vez Paul Graham, seu mentor. Graham reforçou o argumento anos depois: "Sam é extremamente bom[20] em se tornar poderoso."

Samuel Harris Gibstine Altman nasceu em 22 de abril de 1985, filho primogênito de pais judeus, em Chicago, Illinois.

Sua mãe, Connie Gibstine, é médica. O pai dela, Marvin Gibstine, também foi médico, um pediatra que, enquanto fazia parte do Exército dos EUA, foi enviado com a nova esposa para a Alemanha após a Segunda Guerra Mundial. Connie obteve diplomas em Medicina e Direito, desafiando as

normas de gênero de sua geração. Especializou-se em dermatologia, uma profissão com salário estável e horários flexíveis, permitindo que voltasse para casa para cozinhar e estar presente para os filhos.

Antes de entrar na faculdade de Direito da Loyola University Chicago, Connie conheceu Jerold Altman, um colega atraente três anos mais velho. Jerry, filho de um fabricante de calçados[21] e empresário, já havia sido casado uma vez no final da década de 1920, depois de frequentar a Wharton School da Universidade da Pensilvânia e se tornar consultor em Boston. A ex-esposa manteve o nome de solteira. Quando Connie se casou com Jerry, também adotou o sobrenome Altman. Alguns anos mais tarde, o casal se mudou de Chicago de volta para St. Louis, sua cidade natal.

Jerry entrou no ramo imobiliário e de administração de propriedades, chegando a atuar como conselheiro jurídico-chefe e vice-presidente das Roberts Companies, uma imobiliária de St. Louis. Era sociável, apaixonado por habitação acessível e trabalhou em inúmeros projetos comerciais e residenciais voltados ao fomento da comunidade e à revitalização da cidade. Mais tarde, Sam repetiria uma das maiores lições que seu pai lhe ensinou: "Sempre ajude as pessoas e, mesmo que ache que não tem tempo para isso, dê um jeito."

Connie e Jerry tiveram três meninos em sequência: após Sam, nasceram Max e Jack. Cinco anos depois, quando Sam tinha nove, Connie deu à luz Annie, e estava encantada por finalmente ter uma filha. Ela se referia a si mesma como ateia, mas culturalmente judia, ao passo que Jerry era mais religioso. Ele frequentava a sinagoga durante os principais feriados judaicos, como o Pessach, e insistiu para que as quatro crianças tivessem *bat* e *bar mitzvahs*, cerimônias judaicas de maioridade. A racionalidade e disciplina de Connie, somadas à espiritualidade e ao foco em serviço de Jerry, se manifestariam de maneiras distintas em seus filhos.

Desde cedo,[22] Sam era determinado e intensamente curioso. Aos dois anos, aprendeu a usar o videocassete da família; aos três, já o consertava. Quando os pais lhe deram um computador Mac cinco anos depois, o menino rapidamente aprendeu a programar e desmontar a máquina. Adaptou-se bem ao papel de irmão mais velho, às vezes mandando nos mais novos ou cuidando deles. Era extremamente competitivo e sempre insistia em vencer nos jogos de tabuleiro.

Assim como não sabia perder, também tinha um entusiasmo contagiante pela vitória. Quando a avó[23] presenteou cada um dos netos com ações, ele escolheu as da Apple; Jack optou pelas da Applebee's. A discrepância virou piada recorrente na família. Ao longo de vinte anos, as ações de Jack mal cresceram, enquanto as de Sam dispararam. "Suas ações da Apple estão em alta, não quero nem pensar nisso", disse Jack mais tarde, relembrando a história. "Cresceram centenas e centenas de vezes".

"Sim, cresceram muito", respondeu Sam, vangloriando-se.

À medida que Sam crescia, Connie lhe deu uma escolha que ofereceria a

todos os filhos: a possibilidade de transferência para a John Burroughs, uma escola particular local conhecida pelo rigor acadêmico e uma impressionante lista de ilustres ex-alunos. Sam aceitou a proposta; Max também, mas não permaneceu; Jack recusou, e Annie seguiu o irmão mais velho. Em Burroughs, Sam prosperou. Destacou-se acadêmica e socialmente, com sua personalidade extrovertida e humor brincalhão. Interessava-se não apenas por ciências exatas, mas também por escrita e uma série de atividades extracurriculares. Foi editor do anuário, capitão do time de polo aquático e participou da Simulação da ONU, um programa que reúne estudantes em eventos ao redor do mundo para replicar as Nações Unidas e debater políticas públicas. "Lembro-me de pensar — e esta é uma confissão constrangedora: 'Espero que Sam não vá para a área de tecnologia. Ele é tão criativo e um escritor tão bom'", recordou Andy Abbott, seu professor de inglês que, mais tarde, se tornaria diretor da Burroughs. "Eu esperava que ele se tornasse escritor ou algo assim."[24]

Até mesmo naqueles tempos, Altman era carismático e um líder nato. Gostava de testar os limites[25] do que era politicamente aceitável em sua escola mais conservadora. Certa vez, arrumou confusão ao liderar o time de polo aquático em um striptease até ficarem apenas de sunga durante um esquenta, uma assembleia escolar festiva, antes dos jogos. Foi na mesma época em que ele contou aos pais e colegas que era gay. Embora tenha surpreendido a mãe, Connie aceitou a notícia, assim como o restante da família. No entanto, um grupo de estudantes cristãos da escola não reagiu da mesma forma. No Dia Internacional de Sair do Armário, boicotaram uma assembleia que Sam liderava sobre sexualidade. Aos dezessete anos, Altman decidiu confrontá-los em um discurso para todo o corpo estudantil, que, segundo seu orientador universitário, ajudou a expandir a cultura da escola. "Ou você tem tolerância[26] e uma comunidade inclusiva, ou não tem, e não dá para escolher apenas o que convém", diria ele mais tarde, relembrando a frase final.

Por trás da fachada confiante, Sam também era sensível. Preocupava-se com o que as pessoas pensavam sobre ele. Frequentemente, lutava contra a ansiedade, uma característica que o acompanharia na vida adulta. À medida que seu nome crescia[27] no Vale do Silício, às vezes ligava para a mãe com dor de cabeça, convencido de que tinha meningite ou linfoma. Certa vez, ficou tão assustado[28] durante a negociação de um acordo, que precisou deitar-se no chão, sem camisa, com os braços abertos, para tentar se acalmar.

Esses dois lados de sua personalidade — a ambição e a sensibilidade — acabariam moldando sua carreira. Após passar incontáveis horas[29] com Altman para criar seu perfil em 2016, Tad Friend, da *New Yorker*, observou tal dualidade: em qualquer questão, Altman parecia movido tanto por um desejo implacável de avançar quanto por uma atenção constante à necessidade de ter cautela. Ou seja, alcançar a IAG o mais rápido possível, mas sem destruir a humanidade.

Ao se formar em Burroughs, em 2003, Altman deixou o centro-oeste rumo à

Universidade Stanford, atraído por sua proximidade com a indústria de tecnologia. Todavia, não decidiu ingressar no setor imediatamente. Como esperava seu professor Andy Abbott, chegou a considerar ser escritor. Também cogitou, por um curto tempo, seguir carreira como banqueiro de investimento. Por fim, inclinou-se à sua fascinação por programação e computadores. "Percebi que o mundo[30] não precisa do sétimo milionésimo romance nem o valoriza", diria mais tarde. "Não era ali que eu poderia contribuir da melhor forma e, em casos assim, também é geralmente mais difícil ganhar muito dinheiro, ou até mesmo o bastante para viver."

Altman se formou em ciência da computação e desenvolveu um interesse particular por inteligência artificial e segurança. Aprofundava-se nas tarefas;[31] certa vez, segundo um colega, chegou a dissecar um software que deveria usar para um trabalho até chegar à linguagem de programação de baixo nível e acabou encontrando um erro na própria tarefa. No segundo ano, passou a se interessar por tecnologia móvel. Ao descobrir que os telefones celulares[32] logo viriam equipados com GPS, compareceu a um evento de empreendedorismo no campus e subiu ao palco segurando um celular de flip. Fez um convite aberto para que pessoas se juntassem a ele na criação de algo que aproveitasse o recurso de rastreamento de localização.

Na mesma época, conheceu Paul Graham, um empreendedor e influente blogueiro de tecnologia que estava iniciando, com a namorada Jessica Livingston, um novo programa de aceleração de startups chamado Y Combinator. Altman entrou no primeiro grupo de empresas da YC em 2005 como fundador da Loopt, sua nova startup, e passou o verão em Cambridge, Massachusetts, onde o programa começou. A Loopt era uma rede social que usava rastreamento de localização para avisar os usuários quando estavam perto de amigos ou recomendar restaurantes próximos. Ele trabalhou tanto naquele verão e comeu tanto macarrão instantâneo que acabou desenvolvendo escorbuto.

Altman não se arrependeu. "Trabalhem muito[33] no início de suas carreiras", diria mais tarde a jovens fundadores. "Isso rende como juros compostos." Ele nunca retornou a Stanford. No final de 2005, Sam[34] e seus cofundadores já estavam em negociação com as firmas de capital de risco New Enterprise Associates e Sequoia, que lhes ofereceram 5 milhões de dólares em financiamento. Ele decidiu arriscar e abandonou a faculdade.

A Loopt não se tornaria um grande sucesso. Após sete anos de funcionamento,[35] Altman a venderia em 2012 por 43,4 milhões de dólares, aproximadamente o valor que seus investidores haviam aplicado na empresa. Porém, se alguém ouvisse suas entrevistas e os comentários de seus financiadores na época, a startup parecia estar prestes a passar por uma grande transformação.

É mais fácil entender as sementes do sucesso de Altman nessas entrevistas iniciais, quando ele estava vendendo algo muito menos sedutor do que

inteligência artificial; ou seja, um concorrente precoce do Foursquare que acabou não dando certo.

Tanto a habilidade com a mídia quanto o talento para negociações, dois pilares de sua ascensão, baseiam-se em sua excelente capacidade de contar uma boa história. Nesse aspecto, Altman é genial. Mesmo sabendo, ao observá-lo, que sua empresa acabaria fracassando, é impossível não se envolver pelo que ele diz. Fala com uma naturalidade descontraída sobre o posicionamento singular de sua companhia. Sua startup faz parte da grande e irrefreável trajetória da tecnologia. Consumidores e anunciantes estão ávidos pelo serviço. Não aposte contra ele, seu sucesso é inevitável.

"A resposta tem sido incrível",[36] disse ele a Robert Scoble, blogueiro de tecnologia, em junho de 2010, sobre o novo aplicativo de sua empresa, o *Loopt Star*, que permitia a anunciantes oferecerem promoções a usuários, como cupons de restaurantes ou descontos em grupos de varejistas, com base em sua localização. "Ultrapassamos aquele ponto em que o valor percebido ao divulgar onde estou supera minhas preocupações com privacidade ao fazer isso", acrescentou. "Em alguns anos, será normal compartilhar sua localização, e será estranho quando alguém não fizer isso."

"É uma distinção ridícula",[37] disse Altman alguns meses depois à *CNN Business*, sobre a diferença entre a vida online e a presencial; ambas estavam se fundindo com o rastreamento de localização em dispositivos móveis. "O mundo inteiro está se tornando móvel, caminhando para o acesso universal aos seus dados e serviços, onde quer que você esteja", afirmou.

Para Altman, até mesmo discutir os riscos era uma oportunidade de reforçar sua proposta. Quando Jessica Lessin, fundadora do *The Information* e então repórter do *Wall Street Journal*, disse a ele em 2008 que escreveria uma matéria sobre as preocupações em relação à privacidade no rastreamento de localização, ele se ofereceu para ajudar. Enviou-lhe uma longa lista de riscos que a Loopt já havia identificado e suas propostas para resolvê-los. A mensagem implícita era: é assim que o mundo vai funcionar, então é melhor você se preparar. "Ele não queria apenas[38] criar uma startup", escreveu Lessin sobre a experiência. "Queria escrever as regras."

Com a Loopt, Altman construiu as redes e aperfeiçoou as habilidades que se tornariam suas maiores vantagens. Como fundador de startup durante meados dos anos 2000 e início de 2010 na Bay Area, ele se inseriu no centro de uma era de rápido crescimento e novos empreendimentos impressionantes. Convivia com outros empreendedores inquietos, criando conexões primordiais por onde passava. Logo no começo da Loopt,[39] seu escritório ficava no mesmo corredor da nascente startup YouTube. Entre seus *batchmates* da YC — termo utilizado para se referir aos fundadores que participam da mesma turma na YC — estavam Steve Huffman e Chris Slowe, respectivamente cofundador e engenheiro do Reddit. Altman se tornaria membro do conselho do Reddit[40] em 2014, chegando a possuir uma fatia da empresa maior do que a do próprio Huffman. Outro colega de turma era o cofundador da Twitch,

Emmett Shear, que viria a assumir o cargo de CEO interino da OpenAI durante a breve demissão de Altman, quase duas décadas depois.

Altman também aprendeu a melhor forma de apresentar algo à mídia e o caminho mais seguro para fechar acordos extraordinários. Mesmo sendo o CEO de uma startup pouco conhecida, conseguiu negociar parcerias corporativas com as principais operadoras de telefonia móvel dos Estados Unidos. Dizem que o segredo de sua fórmula está na combinação da notável capacidade de ouvir, sua disposição em ajudar e a habilidade de adequar o que tem a oferecer exatamente ao que outro deseja. (Hoje em dia, como um dos superabastados e peças centrais do Vale do Silício, ajuda o fato de ele poder oferecer muito.) É o "Michael Jordan da escuta", dizem alguns; o "Usain Bolt da captação de recursos",[41] segundo Geoff Ralston, que assumiu a direção da YC depois de Altman.

"Essencialmente, ao arrecadar dinheiro de alguém, você está contando uma história sobre o futuro de seu projeto, uma que envolve o projeto em si e a transformação da empresa em um sucesso extraordinário", diz Ralston. "Sam consegue contar uma narrativa da qual você deseja fazer parte, que é envolvente e parece real, até mesmo provável."

Ralston compara isso ao campo de distorção da realidade de Steve Jobs. "Steve conseguia contar uma história que eclipsava qualquer outra parte de sua realidade", diz ele, "fosse porque havia uma distorção da realidade ou porque ela se tornava real. Porque, lembre-se: a questão sobre Steve é que ele realmente criava coisas que mudavam sua realidade. Não era apenas distorção. Era real."

"E, obviamente, Sam também faz isso."

Entretanto, há outro lado nessa história. "Sam se lembra de todos os detalhes sobre você. É extremamente atento. Mas parte disso é que ele usa tais informações para descobrir como influenciá-lo de diferentes maneiras", diz uma pessoa que trabalhou com ele por diversos anos. "Ele é tão bom em se ajustar ao que se diz que você de fato sente que está fazendo progresso com ele. E então, com o tempo, percebe que, na realidade, só está fazendo progresso no mesmo lugar."

Duas vezes durante o período em que Altman comandava a Loopt,[42] líderes seniores da startup procuraram o conselho da empresa e pediram demissão, segundo o *Wall Street Journal*, apresentando duas acusações que o seguiriam até sua breve destituição na OpenAI. A primeira era sobre sua tendência a agir em benefício próprio, e não da empresa, às vezes até mesmo em detrimento dela. A segunda era a respeito de sua aparente compulsão por distorcer a verdade. Esta era mais complicada de definir: de vez em quando, ele mentia sobre detalhes tão insignificantes que era difícil entender por que a desonestidade importava. Com o tempo, porém, esses pequenos "arranhões", como alguém os descreveu, difundiram uma atmosfera de desconfiança e caos na empresa.

De uma forma que viria a definir o restante de sua carreira, Altman saiu

vitorioso da crise. O conselho da Loopt ficou do lado dele.

Apesar do desempenho mediano da startup, Altman também sairia da Loopt em situação muito melhor do que entrou. Usou a empresa como trampolim para se inserir cada vez mais fundo nas redes mais poderosas do Vale do Silício e, então, aproveitou tais conexões para orquestrar uma venda que o deixou rico. Aos 26 anos, lucrou 5 milhões de dólares com a venda da Loopt. Altman se decepcionou — Jobs já valia[43] 256 milhões de dólares aos 25 —, mas logo acumularia muito mais dinheiro. Aos poucos, essa riqueza mudaria seu estilo de vida. Com o tempo, ele deixaria de ir ao mercado, passaria a viajar de jato particular e colecionaria carros esportivos de luxo,[44] incluindo McLarens e um raríssimo Koenigsegg de 5 milhões de dólares, cultivando o apreço por pilotá-los. Por um período, frequentou o Burning Man, o festival anual de arte no deserto, famoso por sua atmosfera psicodélica e sexualmente libertária. Como muitos figurões do Vale do Silício, tornou-se usuário ocasional de cetamina, uma droga recreativa que pode ser prescrita legalmente para aliviar os sintomas de depressão.

Com o sucesso, Altman levou os irmãos consigo. Em 2012, criou a Hydrazine Capital, um fundo pessoal de investimentos, com o irmão Jack, que havia estudado economia em Princeton e estava tentando seguir carreira em *investment banking*. Jack acabaria migrando para o setor de tecnologia e fundaria a Lattice, uma startup que receberia financiamento da YC depois que Sam se tornasse presidente da incubadora. Max, que estudara ciência da computação na Duke University e trabalhara brevemente na Microsoft antes de se tornar operador de mercado, passou a ocupar um posto na Zenefits, outra empresa da YC, em 2014. Durante esse período, os dois irmãos mais novos se mudaram para a casa de Sam, em um arranjo que deveria ser temporário. Os três acabaram morando juntos por muitos anos — um núcleo forte, criado a partir de amor fraternal e laços de negócios.

De importância especial[45] para a formação da carreira de Altman foi sua relação com Paul Graham e Peter Thiel, seus dois mentores.

Conhecido como PG, Graham havia se destacado primeiro como cofundador da startup Viaweb, adquirida pelo Yahoo em 1998 por 49 milhões de dólares, e depois como blogueiro, publicando ensaios populares sobre startups, empreendedorismo e capital de risco. Após fundar a YC, transmitiu suas visões a cada geração de fundadores da incubadora. Toda empresa da YC que se tornava bem-sucedida aumentava o prestígio tanto da incubadora quanto de Graham. Quando Altman vendeu a Loopt, a YC já havia investido em diversas startups que cresceriam ou já estavam prestes a crescer para se tornar empresas bilionárias, incluindo a Dropbox e a Airbnb. A YC tornou-se o clube mais elitista do Vale do Silício. Se você fazia parte dele, ganhava influência instantânea e acesso a maiores recursos, incluindo uma base de clientes formada pelas antigas empresas da YC, investidores mais dispostos a financiá-lo e avaliações mais altas. Caso contrário, a sorte não estaria a seu

lado.

Graham se tornou um influenciador essencial para startups e sua cultura no Vale do Silício. "Muitas pessoas no setor, no ecossistema, passaram a viver e a seguir seus princípios fundamentais sobre o que significava ser um bom fundador e um empreendedor de sucesso", diz Ralston. "Muitos de nós olhávamos para PG em busca de orientação sobre diversas questões." Contudo, Graham também se tornou um alvo frequente de críticas. Defendia a ideia de que a indústria de tecnologia era uma meritocracia, ao mesmo tempo que estruturava a YC como uma fraternidade fechada. Justificava a falta de fundadoras dizendo que a maioria das mulheres não havia sido preparada desde cedo para ter sucesso como empreendedoras da área tecnológica. Após analisar o desempenho dos candidatos em entrevistas na incubadora, identificou trinta ou quarenta fatores, incluindo "um forte sotaque estrangeiro", que indicavam risco de fracasso quando eles manifestavam vários desses fatores ao mesmo tempo. Segundo Graham, isso não era uma falha[46] do sistema de avaliação da YC, mas sim um sinal importante baseado em dados.

O apoio de Graham a Altman foi forte e precoce. Em um post publicado no blog em 2006, contou que conhecera Altman quando ainda era calouro na faculdade. "A Loopt é provavelmente[47] a startup mais promissora de todas as que financiamos até o momento", escreveu Graham. "Mas Sam Altman é um cara muito incomum. Dentro de cerca de três minutos após conhecê-lo, lembro-me de ter pensado: 'Ah, então é assim que Bill Gates deve ter sido aos dezenove anos.'"

Altman rapidamente inspirou Graham[48] a procurar por mais pessoas como ele. Perguntou ao jovem fundador o que a YC deveria colocar em seu formulário de inscrição para descobrir mais pessoas iguais a Sam. A sugestão foi adicionar uma pergunta que Graham logo descreveria como uma das mais importantes: "Por favor, conte sobre a vez que você obteve mais sucesso ao hackear um sistema (não de computador) a seu favor." Ela acabaria encapsulando e incentivando, entre gerações de startups, certa ética de contornar, burlar e quebrar regras com o intuito de dominar.

Quando Altman tinha 23 anos, Graham já o comparava a Jobs. "Sam é, junto com Steve Jobs,[49] o fundador a quem mais me refiro ao aconselhar startups", escreveu. "Em questão de design, pergunto: 'O que Steve faria?'; no entanto, quando o assunto é estratégia ou ambição, pergunto: 'O que Sama faria?'", referindo-se a Altman pelo apelido, que também é seu nome de usuário no X.

Foi a crença singular de Graham em Altman que o catapultou à presidência da YC em 2014, aos 28 anos, dois anos depois de vender a Loopt. Quando Graham perguntou[50] em sua cozinha se ele gostaria de ser seu sucessor, Altman sorriu abertamente. "A YC meio que dita o curso da tecnologia", diria ele mais tarde. "Acho que o objetivo dele é moldar todo o futuro", comentou Graham. A história da sucessão seria repetida tantas vezes

que se tornaria uma lenda do Vale do Silício. “Se Sam sorri, é totalmente intencional”, disse um antigo fundador da YC. “Ele sorriu abertamente apenas uma vez, quando PG disse para assumir a empresa.” A escolha de Graham surpreendeu a muitos, mas ele mantinha convicções firmes. “Não havia uma lista de quem deveria dirigir a YC com Sam em primeiro lugar”, lembraria Livingston. “Era simplesmente ele.”

Peter Thiel se tornou o segundo mentor de Altman. Outro pilar da indústria de tecnologia, Thiel ganhou bilhões ao fundar a PayPal, empresa de pagamentos, e a Palantir, voltada para a mineração de dados, além de ter sido um investidor inicial do Facebook. Como Graham, atraiu sua cota de controvérsias, incluindo o fato de ser explicitamente um raro defensor de Trump entre seus colegas de trabalho durante a eleição de 2016 e financiar secretamente um processo que levaria à queda do Gawker Media, em retaliação por ter sido exposto como homossexual pelo site quase uma década antes.

Após a venda da Loopt, Altman passou por um rompimento com um de seus cofundadores, que também havia sido seu namorado por nove anos. De coração partido e profissionalmente à deriva, tirou um ano sabático, fundou a Hydrazine e levantou 21 milhões de dólares. Thiel, quase vinte anos mais velho, contribuiu com a maior parte do financiamento. Quando Altman se tornou parceiro da YC, usou a Hydrazine para apostar nas empresas do portfólio da incubadora, enquanto também ajudava a Founders Fund, firma de capital de risco de Thiel, a identificar investimentos de alto retorno. A fortuna de Thiel foi multiplicada diversas vezes. Ambos se tornaram extremamente próximos. (A relação entre eles foi descrita como um vínculo[51] com apenas um paralelo: o relacionamento de Thiel como mentor de Mark Zuckerberg, cofundador do Facebook.)

Graham e Thiel influenciaram intensamente a visão de mundo de Altman, sua abordagem para construir negócios competentes e sua habilidade como operador político. Ambos os mentores incutiram nele a importância da escala e a crença na eficiência do capitalismo em detrimento do governo.

“A primeira lição sobre startups[52] que ouvi foi ‘aumentar suas vendas resolverá todos os problemas’”, escreveu Altman em um blog, em 2013,em um post intitulado “Crescimento e governo”, no qual agradecia a Graham e Thiel por moldarem suas ideias. “Isso se revela como outra forma de expressar a crença de Paul Graham de que o crescimento é fundamental.” Para startups, mais vendas eram sinônimo de mais capital, o que levava a melhores talentos e diminuía as tensões internas. Para países, um maior crescimento significava mais inovação tecnológica e, consequentemente, um aumento na qualidade de vida. A disfunção do governo dos Estados Unidos estava ameaçando esse ciclo de crescimento, acrescentou Altman. “Ou você está crescendo, ou está morrendo lentamente”, e o governo norte-americano estava a caminho da morte. “Sem crescimento econômico, a democracia não funciona, porque os eleitores passam a ocupar um sistema de soma zero”, disse.

Essa ideia evoluiria para uma tese central que regularia a carreira e os investimentos de Altman. "A coisa mais importante que as pessoas[53] do setor privado podem fazer para ajudar o país a voltar aos trilhos é restaurar o crescimento econômico", disse ele em 2017. "Nos Estados Unidos, tivemos dois séculos de um crescimento econômico incomparável. Tivemos cem anos de expansão territorial e outros cem anos de nova tecnologia em pleno funcionamento", acrescentou, ignorando a sangrenta história colonial e o complexo passado de exploração de trabalho e meio ambiente da industrialização desenfreada. "As pessoas eram, em geral, muito felizes. E agora não são mais."

"O crescimento econômico sustentável[54] é quase sempre um bem moral", acrescentaria Altman em 2019. "Parte do que me motiva a trabalhar na Y Combinator e na OpenAI é justamente isto: voltar a esse ponto, retomar o crescimento econômico sustentável, retornar a um mundo em que a vida da maioria das pessoas melhora a cada ano e em que sentimos um espírito compartilhado de sucesso."

Quanto à formação de empresas, Altman frequentemente canalizava a estratégia de "monopólio" de Thiel, a crença de que todo fundador deveria "mirar no monopólio" com o objetivo de criar um negócio bem-sucedido. Em 2014, Altman retornou a Stanford, sua *alma mater*, para ministrar uma disciplina intitulada *Como Começar uma Startup*. Convidou Thiel para desenvolver sua filosofia característica em uma palestra chamada "Competição é para perdedores".

Os monopólios são bons,[55] dizia Thiel, porque "são negócios muito mais estáveis e duradouros, você tem mais capital e… isso é um sintoma de que algo realmente valioso foi criado". Construir um monopólio dependia de possuir algum tipo de tecnologia proprietária, efeitos de network, economias de escala e boa reputação de marca. Cada um desses elementos precisava se manter ao longo do tempo. No caso da tecnologia proprietária, era fundamental manter-se na posição de liderança. "Você não quer ser superado por outra pessoa", afirmou Thiel. "Existem muitas áreas de inovação nas quais houve uma imensa quantidade de inovação, mas ninguém saiu lucrando."

Ele citou o exemplo da fabricação de discos rígidos na década de 1980, um setor que apresentava avanços significativos a cada dois anos, mas sempre por empresas diferentes. "Isso trouxe grandes benefícios aos consumidores, mas não ajudou de fato as pessoas que fundaram essas empresas", disse. As companhias precisavam não somente de "um enorme avanço" no começo para estabelecer sua dominância, mas também da garantia de que teriam o "último avanço" para mantê-la, ou seja, "melhorar em um ritmo rápido o bastante para que ninguém jamais conseguisse superá-las".

"Se você conta com uma previsão de futuro em que há muita inovação e outras pessoas criam coisas novas dentro do campo no qual você está trabalhando", concluiu Thiel, "isso é ótimo para a sociedade. Contudo, na realidade, não é tão positivo para o seu negócio."

De ambos os mentores, Altman também aprendeu sobre a importância de construir relacionamentos e criar "efeitos de network" como indivíduo.

"Ouvi diversas[56] teorias diferentes sobre como as coisas são feitas", escreveu em seu blog em 2013. "Aqui vai a melhor: uma combinação de foco e conexões pessoais. Charlie Rose disse isso a Paul Graham, que me repassou a lição." Altman acrescentaria mais tarde um terceiro ingrediente: a autoconfiança. "Para startups, acho[57] muito importante incluir esse traço", disse. "Você realmente precisa acreditar que pode conseguir."

Altman passou a viver seguindo rigorosamente esse mantra. Cultivava relacionamentos com intensidade e disciplina, primeiro oferecendo seu tempo e conselhos táticos e, então, à medida que passou a controlar quantidades crescentes de capital, disponibilizando seu dinheiro. Thiel foi um modelo nesse sentido: seu mentor usava havia muito tempo conselhos e capital para construir a própria rede, e a usava para acumular mais conexões e dinheiro. Para jovens empreendedores e outras pessoas que desejava trazer para seu círculo, Thiel oferecia mentoria e pequenas quantias de capital, assim como acesso. De modo muito similar, Altman aprendeu a usar seus recursos financeiros e sociais de forma estratégica. À medida que o peso de seu nome e sua riqueza cresciam, ele ampliava a abordagem com uma eficiência implacável.

Altman se tornou um anfitrião frequente de jantares e encontros em sua casa para diferentes grupos interligados de pessoas — ligadas ao YC ou às suas empresas, pessoas que compartilhavam seu interesse por investimentos e a ativa e crescente comunidade de empreendedores gays. Transmitia conselhos por mensagens curtas e ligações — de apenas dois minutos — para encaixar mais coisas em sua agenda. Conectava pessoas entre si por e-mail com uma única palavra ("*meet*") ou um único sinal de pontuação ("?"), o que era um hábito famoso de Jeff Bezos, da Amazon, para iniciar uma conversa. Com seu dinheiro, Altman fazia pouquíssimos investimentos de grande porte, optando principalmente por apostas pequenas em alto volume. Com o tempo, de acordo[58] com uma avaliação do *Wall Street Journal* datada de julho de 2024, acumulou vínculos financeiros com mais de quatrocentas empresas por meio da YC, da Hydrazine e de seus outros fundos.

É difícil encontrar pessoas no círculo íntimo de Altman que não tenham algum tipo de relação financeira com ele. Seu segundo investimento em startups e o mais bem-sucedido foi na Stripe, empresa de tecnologia de pagamentos apoiada pela YC, cujo executivo principal de tecnologia era Greg Brockman. Altman investiu cedo no Airbnb, outra startup apoiada pela YC, cujo cofundador e CEO, Brian Chesky, é um de seus confidentes mais próximos. Ele investiu na Conception, empresa de biotecnologia de seu ex-namorado e amigo Matt Krisiloff. Investiu em negócios com Lachy Groom, outro ex-namorado e proeminente investidor de risco independente. Para essas pessoas, isso foi um testemunho da generosidade de Altman. Com frequência, ele oferece seus recursos, seja abrindo suas casas para que elas se hospedem,

seja apoiando-as financeiramente. Chegou a se esforçar para ajudar até completos desconhecidos, enviando fundos a um homem na Etiópia que lhe escreveu pedindo ajuda para comprar um laptop, disse alguém. Durante a crise do Banco do Vale do Silício em 2023,[59] quando uma corrida a uma instituição financeira essencial para startups da área levou à maior falência bancária desde 2008, Altman enviou dinheiro sem qualquer documentação para empresas a fim de salvá-las de fecharem as portas ou demitirem seus funcionários, lembra Krisiloff. "É uma característica extremamente rara", diz Groom, "e de fato me contagiou — a generosidade. Sinto-me muito grato por isso."

Seguindo outra lição de Thiel, Altman desenvolveu a mesma abordagem com políticos. Contudo, enquanto o mentor utilizava sua riqueza para apoiar candidatos republicanos, investindo dezenas de milhões em suas campanhas, ele se envolveu cada vez mais na política na direção oposta, organizando eventos de arrecadação e assinando cheques para democratas. Por um tempo, as diferenças políticas entre Thiel e Altman foram motivo de tensão na relação dos dois. Em 2017, Altman aprofundou essas discordâncias. Ele fez uma turnê pelos Estados Unidos, assim como Zuckerberg, outro pupilo de Thiel, e conversou com cem apoiadores de Trump. Também cogitou entrar na política pessoalmente, concorrendo ao governo da Califórnia, argumentando que isso o colocaria à frente da quinta maior economia do mundo, um forte trampolim para reparar o que via como disfuncional no sistema político. Publicou um manifesto[60] chamado *The United Slate* [A chapa unida], com três princípios: (1) prosperidade a partir da tecnologia; (2) justiça econômica; e (3) liberdade pessoal. Em seguida, organizou grupos focais para testar sua candidatura. Pessoas próximas brincavam que ele deveria mirar a presidência.

No fim, Altman nunca embarcou na política — os grupos focais acharam que ele parecia jovem demais —, mas começou a agir como um político. Nos primeiros anos à frente da YC, ainda tinha bochechas infantis, possuía apenas um paletó e se sentava com uma perna levantada ou empoleirado como um pássaro sobre a cadeira. Às vezes, falava de forma irreverente e com hipérboles casuais, pontuando as frases com palavrões. Tornou-se mais impaciente, rapidamente demonstrando irritação com pessoas ineficazes, chegando a programar um software para avaliar fundadores da YC com base no tempo que levavam para responder seus e-mails.

Alguns anos depois, havia refinado a aparência e suavizado seus extremos. Trocara camisetas e shorts cargo por camisas Henley ajustadas e jeans, ganhara 8 quilos[61] de músculo em apenas um ano para preencher o corpo franzino e aprendera a falar menos, perguntar mais e projetar uma modéstia pensativa ao franzir o cenho. Em ambientes privados e com amigos próximos, ainda demonstrava lampejos de raiva e frustração; já em público e com conhecidos, fazia o tipo bom moço. Dava crédito às pessoas facilmente e enviava mensagens em letras minúsculas cheias de carinhas felizes e tristes. Passava seu número pessoal aos funcionários, incentivando-os a ligar a qualquer

instante e respondendo aos feedbacks com atenção impressionante. Evitava expressar emoções negativas, envolver-se em confrontos e dizer "não" às pessoas. Certa vez, quando a OpenAI demitiu um funcionário, entrou em contato pessoalmente para oferecer cetamina e bebidas como consolo. "Acho que todos os relacionamentos de Sam acabam de forma positiva, queira você ou não", disse o funcionário.

Altman se tornou uma instituição por si só. A YC era sua plataforma e seu combustível. Ele converteu o poder e a rede de contatos da empresa em seu próprio poder e sua própria rede de contatos. Tais conexões pessoais e a reputação pública tornaram-se sua maior moeda. Altman se encontrava regularmente com formuladores de políticas, que o viam como uma porta de entrada para o Vale do Silício. Em 2016, foi Ashton Carter,[62] secretário de defesa de Obama, quem buscou o conselho de Altman sobre como sua agência poderia acessar o reservatório de jovens talentos em tecnologia. Três anos depois,[63] no dia em que Altman deixou a YC, em março de 2019, foi a vez de Chuck Schumer. Na época, líder da minoria no Senado dos EUA, Schumer fez uma visita clandestina à OpenAI com sua equipe do Serviço Secreto. "Vocês estão realizando um trabalho importante", disse aos funcionários no escritório, enquanto se sentava lado a lado com Altman em poltronas diante de uma TV que projetava uma lareira acesa. "Não o entendemos totalmente, mas é importante", acrescentou Schumer. "E conheço Sam. Vocês estão em boas mãos."

A ascensão de Altman[64] também viria a um custo crescente, à medida que ele acumulava mais e mais difamadores e inimigos declarados, que repetiam as acusações de seus principais subordinados na Loopt: a busca egoísta por poder e a desonestidade compulsiva. Enquanto muitas pessoas que se beneficiaram das palavras, da riqueza e das redes de Altman se tornaram aliadas leais e inabaláveis, outras começaram a vê-lo como perversamente capaz de manipular situações ao próprio favor. Para alguns, incluindo seus parceiros na YC e outros indivíduos influentes, isso podia ser um incômodo. Para funcionários e pessoas com níveis de influência muito menores, podia ser uma fonte de medo. Para outros mais, que discordavam veementemente de sua visão de mundo, ele representava uma enorme ameaça.

Para a aflição e, mais tarde, também ira de Altman, sua ascensão arruinaria seu relacionamento com a irmã. Quando crianças e até os vinte e poucos anos, Sam e Annie eram próximos. Ela, a mais nova; ele, o mais velho, seu protetor. Annie tinha uma inclinação para a ciência e as artes, sendo a mais expressiva emocionalmente. Às vezes, ele gostava de pedir a opinião dela sobre parceiros românticos, de compartilhar com a irmã suas preocupações e sentimentos. Porém, ao passo que Sam imergia cada vez mais no Vale do Silício, Annie o via levantar barreiras mais e mais espessas ao redor de seu lado mais sensível. Ele contava sobre novas táticas psicológicas que havia aprendido, lembra ela, como usar menos palavras em um e-mail para parecer

mais poderoso como líder empresarial.

A princípio, a possibilidade de aquele lado sensível não existir mais a entristeceu e, depois, a assustou. “Eu definitivamente ainda tinha vislumbres dele por um tempo, e por isso continuei próxima”, disse ela. “Mas então comecei a ser a pessoa prejudicada por ele.”

Quando Sam enriqueceu, segundo ela, seu então namorado criou uma regra: para cada item de alto valor que Sam comprasse, ele precisava doar a mesma quantia para uma causa beneficente. Por um tempo, isso funcionou como um freio para o rápido crescimento de seu estilo de vida; no entanto, à medida que ganhava mais dinheiro do que podia gastar, Annie percebeu que o relacionamento do irmão com a fortuna se complicava. Na visão dela, Sam começou apenas a acumular enquanto se afastava progressivamente das pessoas necessitadas. Diversas vezes, entre o final de 2019 e a primeira metade de 2020, ele e o restante da família recusaram ou relutaram em fornecer à Annie o que ela via como suporte financeiro emergencial para ajudar a pagar o aluguel e despesas médicas, de acordo com extensa correspondência compartilhada comigo. Na época, ela enfrentava problemas seríssimos de saúde física e mental, exacerbados pelo falecimento repentino do pai. Isso causou instabilidade quanto à moradia e, desesperada para sobreviver, recorreu ao trabalho sexual para ganhar dinheiro. No verão de 2020, enquanto a OpenAI começava a ganhar sua primeira grande onda de atenção pública sob a liderança de Sam, Annie cortou contato com a família.

Há argumentos de que Sam, assim como os irmãos, seguia o exemplo da mãe na tentativa de forçar a irmã a alcançar a independência financeira. É uma história familiar complicada e dolorosa, difícil de julgar com base em informações parciais. Em uma declaração pública[65] realizada em janeiro de 2025, Sam, a mãe e os dois irmãos expressaram seu amor e a preocupação por Annie e classificaram todas as acusações como “totalmente falsas.” Em resposta a meus pedidos[66] de entrevista e solicitações detalhadas em busca de comentários, Connie Gibstine forneceu uma visão mais curta de uma declaração semelhante e negou-se a elaborar mais; Sam, por meio da equipe de comunicações da OpenAI, e seus irmãos não responderam.

Ainda assim, a experiência de Annie contém paralelos impressionantes com muitos dos temas explorados nestas páginas: o abismo cada vez maior entre aqueles que se beneficiam e os que ficam para trás na suposta marcha do progresso; a perda de livre-arbítrio e voz entre os desfavorecidos perante esse fosso que cresce rapidamente; os limites da cessão de tanto poder não apenas às empresas, mas aos indivíduos que as comandam, sem a estrutura necessária para prover contrapesos proporcionais. As ações de Annie também tornariam sua história uma parte inescapável para compreender a trajetória da OpenAI e seu impacto no desenvolvimento da IA: em 2021, ela decidiu tornar públicas sérias acusações contra Sam, afirmando que ele havia abusado sexualmente dela quando criança — que sua família chamou de “a pior” de todas as suas “falsas” acusações — e que ele e o restante dos parentes a abandonaram

quando estava em seu pior período de vulnerabilidade. Annie entraria com um processo[67] contra Sam pelo suposto abuso em 6 de janeiro de 2025, dois dias antes do trigésimo primeiro aniversário dela, para cumprir o prazo de prescrição para tais casos em Missouri. Os esforços persistentes da irmã de Altman para apresentar suas acusações e contar seu lado da história afetariam Sam e influenciariam outros executivos da OpenAI enquanto lidavam com o crescente impacto global e a proeminência dele e da empresa.

Cada uma dessas peças do quebra-cabeça — a ascensão de Sam, seu caráter e relacionamentos, a segmentação que deixava em seu rastro, os fluxos de dinheiro e poder — aponta para o caminho que levou à sua súbita e curta destituição. Por um breve momento, o restante do mundo entreviu as lutas que aconteciam nos mais altos escalões para ditar o futuro da inteligência artificial. Isso revelaria o quanto a busca pela dominação dessa tecnologia, que já está reestruturando a sociedade e transformando o mundo, repousa, em última análise, sobre os valores polarizados, os egos conflitantes e a humanidade complicada de um pequeno grupo de pessoas falíveis.

CAPÍTULO 2

# UMA MISSÃO CIVILIZADORA

Greg Brockman foi o primeiro a se comprometer a criar a OpenAI. Para ser cofundador de Brockman, Altman escolheu pessoalmente Ilya Sutskever, então pesquisador de IA no Google, a quem enviou um e-mail sem explicações para convidá-lo ao jantar no Rosewood no verão de 2015. Sutskever, animado, aceitou ao saber que Musk estaria presente.

Brockman e Sutskever formavam uma dupla curiosa. Alto e robusto, de comportamento amigável, Brockman era engenheiro e tinha o típico perfil para startups, assim como Altman. Cresceu[1] em uma fazenda em Dakota do Norte. Entre ordenhas de vacas, apaixonou-se por matemática e, depois, por ciência. Em 2008, matriculou-se em Harvard e transferiu-se para o MIT dois anos mais tarde. Após mais um semestre, desistiu totalmente da faculdade, incapaz de suportar mais estudos quando poderia estar no mundo real desenvolvendo produtos. Mudou-se para a Bay Area e entrou na Stripe, na época uma startup com apenas três outras pessoas; impressionou tanto os fundadores com seu talento em codificação que se tornou diretor de tecnologia. Em cinco anos, criou o protótipo de muitos dos primeiros produtos da empresa, contribuindo com seu crescimento até ela se tornar uma potência fintech que fornece infraestrutura de pagamentos digitais para empresas como a Amazon e a Shopify. A trajetória lhe proporcionou uma riqueza significativa e o raro status no Vale do Silício de ter ajudado a criar uma empresa multibilionária.

Sutskever era o cientista. Magro e esguio,[2] nasceu na União Soviética e foi criado em Israel, onde se distinguiu como um prodígio da matemática. Depois de terem dificuldade para encontrar professores que acompanhassem o avanço de Sutskever, seus pais o matricularam, ainda na oitava série, em cursos da Open University de Israel. Aos dezesseis anos, mudou-se para Toronto e cursou o ensino médio por apenas um mês antes de ser admitido na Universidade de Toronto, em 2003, como aluno do terceiro ano de graduação. Foi lá que Sutskever conheceu Geoffrey Hinton, professor britânico-canadense que havia realizado trabalhos seminais em pesquisa de inteligência artificial. Hinton se tornaria a única pessoa a quem Sutskever chamaria de

mentor e teria profunda influência em sua carreira e vida. Em 2012, com Alex Krizhevsky, outro aluno de doutorado de Hinton, impressionou o mundo da IA ao dominar uma competição acadêmica chamada ImageNet, que desafiava equipes a criarem softwares capazes de identificar automaticamente objetos em fotos. Enquanto as demais equipes lutavam[3] para reduzir a taxa de erro de seus programas para menos de 25%, Hinton, Sutskever e Krizhevsky a derrubaram para 15%. No início do ano seguinte, o Google anunciou a aquisição da DNNresearch, empresa recém-criada pelos três, em um leilão acirrado de 44 milhões de dólares. A operação transformou os acadêmicos em multimilionários e desencadeou a primeira grande corrida pela comercialização da inteligência artificial. "A impressão era a de que estávamos[4] dentro de um filme", lembrou Hinton.

Brockman e Sutskever se encontraram pela primeira vez durante o jantar no Rosewood. Assim como eles, os demais convidados eram empreendedores ou cientistas. A conversa alternava entre deliberações acadêmicas sobre diferentes abordagens à pesquisa em IA e a fixação de Musk em saber se ainda havia tempo para ultrapassar a DeepMind e o Google, algo que todos consideravam essencial para corrigir o rumo do desenvolvimento da inteligência artificial. Todos concordavam que a limitação crítica era o talento: a maioria dos melhores pesquisadores de inteligência artificial, como o próprio Sutskever, estava empregada, se não pelo Google, então por outras gigantes da tecnologia, com salários, benefícios e estabilidade extravagantes.

A IAG também ocupou o centro das discussões, o que era bastante incomum na época. A maioria dos cientistas sérios considerava a ideia de replicar digitalmente uma inteligência que equivalesse de fato à humana como ficção científica ou, no mínimo, algo distante por muitas décadas. Declarações ousadas de que estaria ao alcance, a ponto de justificar investimentos imediatos, eram vistas em grande parte como pseudociência e charlatanismo. Ainda assim, Hassabis havia adotado esse termo para descrever as ambições da DeepMind, apesar da crença, entre seus próprios pesquisadores, de que aquilo soava como marketing de mau gosto e sem pudor. O grupo reunido no Rosewood acreditava igualmente que o mesmo objetivo — a IAG — seria a melhor forma de definir as próprias aspirações, caso de fato pretendessem formar um concorrente à altura da organização de Hassabis.

Mesmo para Sutskever, que secretamente[5] acreditava que a IAG era possível e mais tarde defenderia com unhas e dentes algumas das previsões mais agressivas sobre a velocidade de sua criação, aquele tipo de conversa era imprudente e o deixava um tanto desconfortável. Temia que, se outros pesquisadores descobrissem que ele estava discutindo publicamente a busca por esse objetivo, pudesse perder a credibilidade dentro da comunidade científica. Todavia, tais preocupações não afetavam Brockman, um forasteiro no campo, sincero em sua convicção de que a IAG, com dedicação e foco suficientes, poderia estar logo ali, ao virar da esquina. O fato de Sutskever e os demais pesquisadores estarem dispostos ao menos a considerar em particular a

viabilidade de um novo laboratório apenas poderia ter reforçado a confiança de Brockman. Naquela noite, quando Altman o levou de volta do hotel a São Francisco, Brockman lhe disse que estava pronto para se dedicar ao projeto.

No Vale do Silício, há um ditado popular: tornar-se cofundador é como entrar em um casamento. Assim como em qualquer parceria em que há comprometimento, Brockman e Sutskever cortejaram um ao outro depois que Altman diplomaticamente disse ao primeiro que ele precisava ser acompanhado por alguém que entendesse de pesquisa em IA. Algumas semanas depois do jantar no Rosewood, os dois marcaram outra refeição sozinhos em Mountain View. Foi uma combinação perfeita. "Eu sabia que ia dar certo,[6] mesmo tendo acabado de conhecê-lo… Ilya e eu tivemos uma interação extremamente intensa", escreveu Brockman mais tarde em seu blog. "Nossas ideias se ampliaram e se complementaram."

Nos meses seguintes, enquanto Sutskever refletia sobre a proposta, Brockman assumiu a missão de convencer outros a se juntarem à nova e ambiciosa empreitada. Ligava para figuras de destaque na área para pedir recomendações de talentos de ponta em IA; jantava com professores universitários para inquirir sobre seus melhores alunos; pesquisava profundamente cada candidato antes de qualquer conversa, a fim de recrutá-los de modo mais persuasivo. Mais tarde, Altman exaltaria[7] esses esforços iniciais em um post intitulado apenas "Greg" em seu blog. "Muita gente me pergunta como é o cofundador ideal", escreveu. "Agora eu tenho uma resposta: Greg Brockman."

Altman e Musk também tiveram sua parcela de conversas sobre recrutamento, aos poucos convencendo pesquisadores que resistiam à ideia da IAG. "A IAG pode estar longe,[8] mas e se não estiver?", lembrou Pieter Abbeel, professor da Universidade da Califórnia, Berkeley, sobre o que Musk lhe perguntou. "E se houver apenas 1% ou 0,1% de chance de isso acontecer nos próximos cinco a dez anos? Não deveríamos pensar com muito cuidado sobre o assunto?" Abbeel acabaria se juntando à OpenAI como consultor de pesquisa e, mais adiante, em tempo integral, com vários de seus alunos de doutorado.

No início, muitas das pessoas que Brockman abordou estavam dispostas a se juntar ao projeto apenas se outras também o fizessem. Sem se deixar abater, ele convidou seus dez[9] engenheiros e pesquisadores mais desejados para discutir suas hesitações e reacender o entusiasmo durante um encontro regado a vinho no Vale do Napa. Fretou um ônibus para levar e trazer todos, para que pudesse continuar falando de sua proposta durante a viagem de mais de uma hora na ida e na volta. Três semanas depois, perto do prazo estabelecido por Brockman, quase todos haviam aceitado a oferta.

Enquanto Musk, Altman e Brockman discutiam a melhor forma de posicionar a OpenAI no lançamento, todos estavam plenamente cientes da importância da percepção pública sobre a empresa. Concordaram com a

proposta de Altman de torná-la uma organização sem fins lucrativos e de enfatizar a abertura que o próprio nome evocava. A OpenAI, o "anti-Google", conduziria suas pesquisas em prol de todos, revelaria os códigos de sua ciência e seria modelo de transparência.

"Espero que entremos[10] neste campo como um grupo neutro, visando colaborar amplamente e mudar o diálogo para que seja sobre a vitória da humanidade, e não de um grupo ou empresa em particular", escreveu Brockman a Musk e Altman em novembro de 2015. "(Acho que essa é a melhor forma de nos lançarmos como uma instituição de pesquisa de ponta.)"

"Há muito valor em fazer o público torcer para que tenhamos sucesso", respondeu Musk mais tarde naquele mês, sugerindo reescrever o anúncio da formação da OpenAI para ter um apelo mais amplo. "Precisamos usar um número muito maior do que 100 milhões de dólares para não parecer incompetência em comparação com o que o Google ou o Facebook estão gastando", acrescentou. "Acho que devemos dizer que estamos dando início a um compromisso de financiamento de 1 bilhão. O valor é real. Cobrirei qualquer quantia que os outros não disponibilizarem."

Em correspondências posteriores, o grupo reconheceu que poderia voltar atrás em seus compromissos com a transparência depois que a narrativa cumprisse seu propósito e conforme surgisse a necessidade, por exemplo, para evitar que agentes mal-intencionados tivessem acesso à tecnologia. "À medida que nos aproximarmos de criar uma inteligência artificial, fará sentido começarmos a ser menos abertos", observou Sutskever ao trio em janeiro de 2016, logo após o lançamento da OpenAI. "O 'Open' do nome significa que todos devem se beneficiar dos frutos da IA depois de sua criação, mas não tem problema algum não compartilhar a ciência por trás disso." "Certo", respondeu Musk.

Em dezembro de 2015, o anúncio foi divulgado na noite de uma sexta-feira, para coincidir com a conferência Neural Information Processing Systems, o maior evento anual de pesquisa em IA, onde Hinton e Sutskever haviam leiloado a DNNresearch três anos antes. A postagem no blog, intitulada "Apresentando a OpenAI", listava cada um dos nove membros fundadores, incluindo Brockman, que seria o diretor de tecnologia (CTO), e Sutskever, como diretor de pesquisa. Musk e Altman seriam copresidentes. Altman e Brockman também se juntaram a Musk em seu compromisso de garantir que o laboratório recebesse um bilhão de dólares em financiamento. Jessica Livingston, Peter Thiel e Reid Hoffman, cofundador do LinkedIn, seguiram o mesmo caminho. Hoffman havia trabalhado com Musk e Thiel no PayPal e frequentemente investia com eles em startups como parte da "máfia do PayPal". "Esperamos gastar apenas uma pequena fração disso nos próximos anos", dizia a publicação sobre o financiamento.

Nos momentos finais antes do anúncio, Sutskever quase permaneceu no Google. A todos os outros membros fundadores,[11] a OpenAI havia oferecido um salário-base de 175 mil dólares e ações da Y Combinator ou da SpaceX.

No entanto, a Sutskever[12] o laboratório ofereceu quase 2 milhões, uma soma exorbitante para uma organização sem fins lucrativos. Ainda assim, o Google ofereceu[13] mais, e depois ainda mais, duplicando ou triplicando o valor, na tentativa de mantê-lo. Musk e Altman adiaram[14] o anúncio da empresa repetidas vezes enquanto Sutskever agonizava sobre a decisão, ligando para os pais e recebendo apelos de Musk e Brockman. No fim, a oferta vertiginosa do Google reforçou a ele o motivo pelo qual uma organização sem fins lucrativos como a OpenAI é tão necessária.

Naquele dia, Musk marcou a ocasião com um e-mail à equipe fundadora, comprometendo-se solenemente com o sucesso da OpenAI. "Nossa consideração mais importante é o recrutamento dos melhores profissionais", escreveu, prometendo apoiar e fazer o possível para ser útil. "Estamos em desvantagem, em termos de números e recursos, em relação a organizações que vocês conhecem muito bem", acrescentou, "mas contamos com a razão ao nosso lado, e isso vale muito. Gosto de nossas chances." Para evitar que outras ofertas[15] tentassem atrair integrantes da equipe fundadora, a OpenAI imediatamente aumentou o salário-base de todos em mais 100 mil dólares.

Mais adiante, Musk relataria ter enfrentado[16] a fúria de Larry Page por ter pessoalmente "roubado" Sutskever. Os dois quase não voltaram a se falar, uma vez que suas visões sobre o desenvolvimento da IA continuaram a divergir. No entanto, o recrutamento da OpenAI ficou repentinamente mais fácil. Projetos ambiciosos e associações com bilionários eram um poderoso atrativo no Vale do Silício e, em poucos meses, o número de funcionários dobrou.

A falta de clareza, o cheque milionário e a adoração a bilionários representavam o Vale do Silício em seu auge eufórico de poder irrestrito. Contudo, a OpenAI havia sido inteligente em seu posicionamento: equilibrava-se na fronteira entre a versão tecno-chauvinista do local e uma vertente mais consciente que começava a emergir. No ano seguinte, a ascensão espetacular de Donald Trump e sua vitória na eleição presidencial dos Estados Unidos em 2016 chocariam a força de trabalho de inclinação progressista da indústria tecnológica, levando a um período de autorreflexão. À medida que a agitação tomava conta de empresas como a Meta e o Google e o sentimento de "revolta contra a tecnologia" dominava o público, pesquisadores de IA também começaram a questionar se o campo não havia avançado rápido demais ao atrelar suas tecnologias a interesses corporativos.

Uma análise dos impactos sociais da comercialização da pesquisa em inteligência artificial trouxe um balanço perturbador: softwares automatizados vendidos[17] para a polícia, corretores de hipoteca e credores estavam acentuando a discriminação racial, de gênero e de classe. Algoritmos que alimentavam o feed de notícias do Facebook e o sistema de recomendações do YouTube provavelmente haviam polarizado o público, alimentando desinformação e extremismo, facilitando interferência eleitoral e, no caso mais

horrendo, o do Facebook, incitando uma limpeza étnica[18] em Mianmar.

Entretanto, aceitar dinheiro do governo, que era a principal alternativa de financiamento, trazia as próprias armadilhas éticas. Em 2018, milhares de funcionários do Google protestaram contra um contrato secreto da empresa com o Pentágono, referente ao programa conhecido como Projeto Maven, voltado ao desenvolvimento de drones de vigilância movidos por IA. Os funcionários alertaram que as capacidades[19] desenvolvidas poderiam abrir caminho para armas autônomas; o Pentágono, que afirmou não ter tal intenção, mudaria de posição mais tarde com a guerra da Ucrânia.

Entre os pesquisadores de inteligência artificial, surgiu um bordão cínico: vender-se à big tech ou ao complexo militar-industrial ou então abandonar as pesquisas. Entre esses dois extremos, a OpenAI parecia representar uma terceira via, não corrompida pelo lucro ou pelo poder do Estado. "Era um farol[20] de esperança", disse Chip Huyen, engenheira de machine learning, ou aprendizagem de máquina, e blogueira popular no meio tecnológico, observando de fora.

Nem todos se impressionaram. Na noite do lançamento da OpenAI, em dezembro de 2015, Timnit Gebru, a pesquisadora de IA que questionara Musk sobre priorizar as ameaças da inteligência artificial em detrimento das mudanças climáticas, não conseguia acreditar no anúncio.

Durante toda aquela semana,[21] a estudante de pós-graduação da Stanford, refugiada eritreia nascida na Etiópia que se mudara para os Estados Unidos na adolescência, vinha sendo lembrada do alto custo de ser uma mulher negra em um ambiente dominado por homens brancos. Era sua primeira vez participando da multidão de pesquisadores de IA na conferência anual Neural Information Processing Systems, então chamada de NIPS e posteriormente renomeada como NeurIPS. Gebru era uma das únicas pessoas negras presentes. No ano seguinte, na conferência de 2016, ela quantificaria isso de forma literal, contando apenas outras seis pessoas negras entre os 8.500 participantes. Em Stanford, costumava brincar sobre a falta de pesquisadores negros no campus, dizendo que poderia fundar um grupo chamado Negros na IA e realizar as reuniões sozinha. Chegou a imaginar criar um canal no YouTube para satirizar a situação, mudando o penteado e encenando personalidades diferentes para dramatizar suas conversas consigo mesma. No entanto, por mais que tentasse encarar seu isolamento com humor, a conferência de 2015 a lembrou de como as coisas podiam rapidamente se tornar hostis. Certa noite, em uma festa, ela foi buscar um copo d'água quando um grupo de rapazes bêbados usando camisetas da Google Research percebeu sua presença e decidiu fazer dela o alvo de sua diversão. Cercaram-na. Um deles a forçou a abraçá-lo, enquanto outro lhe deu um beijo na bochecha ao tirar uma foto humilhante. Na mesma conferência, uma amiga de Gebru foi assediada por um professor.

Naquele momento, havia ali um grupo de pessoas — nove dos onze eram homens brancos — recebendo quantias até então inéditas, discorrendo sobre a

possibilidade teórica de uma superinteligência malévola dominar o mundo e propondo combatê-la por meio da construção de uma superinteligência melhor.

Naquela noite, Gebru redigiu uma crítica feroz ao que havia presenciado, em uma carta aberta anônima: o espetáculo, a exaltação quase sectária das celebridades do mundo da IA e, acima de tudo, a homogeneidade esmagadora das pessoas encarregadas de construir e moldar uma tecnologia de tamanha consequência. Essa cultura uniforme não apenas afastava pesquisadores talentosos, mas também levava a uma concepção perigosamente estreita da inteligência artificial e de quem poderia se beneficiar dela.

"Não precisamos projetar o futuro para ver os possíveis efeitos adversos da IA", escreveu Gebru. "Eles já estão acontecendo."

No voo de volta para casa depois da conferência, ela reconsiderou a ideia de publicar a carta anonimamente. Em vez disso, postou uma versão mais curta e moderada de sua crítica no Facebook, usando o próprio nome.

Algumas semanas depois, Gebru escreveu um e-mail com o assunto "Timnit diz olá".[22]

"Quando compareço a conferências de visão computacional, costumo ser a única pessoa negra presente", escreveu. "Mas agora vi cinco de vocês :) e pensei que seria legal se criássemos um grupo chamado Negros na IA, ou pelo menos soubéssemos da existência uns dos outros."

Um a um, ela adicionou os e-mails dos pesquisadores. E então clicou em "enviar".

Nos primeiros dias da OpenAI, Altman e Musk mal apareciam em seus postos de copresidentes. Ocupados com a ampla lista de outros empreendimentos, deixaram Brockman e Sutskever encarregados de levantar a organização. Enquanto Sutskever reunia pesquisadores para que compartilhassem suas melhores ideias, Brockman se dedicava intensamente a desenvolver a cultura organizacional certa.

Alguns anos depois, Brockman[23] me contaria sobre seu modo de pensar. Para se preparar, leu todos os livros que conseguiu encontrar sobre ambiciosas empreitadas científicas e tecnológicas da história dos Estados Unidos: a construção da ferrovia transcontinental, a lâmpada de Thomas Edison e a rede inicial de computadores que lançaria as bases da internet moderna. Ele os absorvia como textos religiosos, buscando pistas e orientações sobre como estruturar o próprio empreendimento.

Uma história que lhe era particularmente cara era a provavelmente duvidosa anedota em que John F. Kennedy se aproxima de um zelador que segura uma vassoura no centro espacial da NASA. "Kennedy pergunta a ele: 'Senhor, o que você está fazendo?' E ele responde: 'Ah, estou ajudando a enviar um homem para a Lua'", contou Brockman, claramente maravilhado. "Que todos tenham esse senso de missão e propósito… Acho que é algo realmente incrível, e que não vejo refletido no que normalmente acontece hoje em dia."

Mais tarde, Brockman acrescentou: “Sinto realmente que nós, americanos, paramos de ousar sonhar.”

Para ter sucesso, ele acreditava que a OpenAI precisava desse mesmo nível de alinhamento; cada pessoa, em todos os níveis da empresa, precisava ser como aquele zelador. Contou-me que, de fato, nos primeiros meses da OpenAI, quando todos trabalhavam de seu apartamento, ele incorporava esse espírito ao pé da letra e passava boa parte do tempo lavando os copos dos funcionários. Criou uma política que exigia que todos trabalhassem no escritório de São Francisco, uma regra que a empresa manteria até a pandemia. Tal exigência, é claro, trazia alguns custos; nem todos queriam morar na Bay Area, ele admitiu. Em outras entrevistas, descobri que isso era particularmente verdadeiro para mulheres e pessoas não brancas que, como Gebru, sentiam-se alienadas pela cultura branca e masculina predominante na indústria tecnológica. Para Brockman, porém, a coesão era mais importante, e estar fisicamente juntos favorecia o ocasional intercâmbio de ideias.

Brockman decidiu também chamar todos os funcionários da OpenAI de “membros da equipe técnica”, inspirado no Xerox PARC, o lendário laboratório de pesquisa e desenvolvimento em Palo Alto, que adotara o mesmo termo seguindo a tradição do igualmente renomado Bell Labs, em Nova Jersey, a fim de criar um ambiente de trabalho mais democrático.

Ao considerar as críticas dirigidas à OpenAI devido a sua busca pela IAG, Brockman traçava paralelos com a invenção da lâmpada de Edison. “Um comitê de especialistas ilustres disse: ‘Isso nunca vai funcionar’, e um ano depois, ele entregou o produto.” “Como isso foi *possível*?”[24] Foi, como escreveu o autor de livros científicos Arthur C. Clarke em *Perfil do futuro*, “uma falha de imaginação”.[25]

Entre os presentes no jantar no Rosewood estava Dario Amodei, um neurocientista computacional que se tornou pesquisador de IA. Na época, estava trabalhando no laboratório de inteligência artificial da empresa chinesa Baidu, com sede no Vale do Silício, antes de passar brevemente pelo Google. Daniela Amodei, sua irmã, trabalhara com Brockman na Stripe e, quando Brockman começou a se envolver seriamente com o desenvolvimento de IA, recorreu a Dario em busca de recursos de aprendizado. Dario não se juntou à OpenAI de imediato, mas ficou intrigado com a proposta. Sob a influência de Musk, a empresa parecia se destacar dos demais laboratórios por ser a mais disposta a focar a suposta segurança da IA.

Em 2016, ainda no Google, Amodei foi coautor de um artigo fundamental para a disciplina, articulando como um problema central na segurança da IA “o contratempo de acidentes[26] em sistemas de aprendizagem de máquina, definido como o comportamento não intencional e prejudicial que pode surgir de um projeto deficiente de sistemas de IA no mundo real”. Isso se diferenciava de outros desafios relacionados à inteligência artificial, escreveram ele e seus coautores, incluindo privacidade, segurança, justiça e

impacto econômico. Nesse contexto, "segurança" significava evitar a IA descontrolada e desalinhada — a raiz do que permitiria, conforme descrito por Nick Bostrom, que a superinteligência se tornasse uma ameaça existencial.

Para Amodei, não havia outro assunto mais importante a se trabalhar do que evitar que uma IA super-humana causasse resultados catastróficos, inclusive a extinção do homem. Ambos os irmãos Amodei eram simpatizantes do movimento do altruísmo eficaz, ou AE, uma ideologia controversa que surgiu entre filósofos na Universidade de Oxford, onde Bostrom estava baseado, e que se consolidou no Vale do Silício. Com o tempo, o movimento, que prega dedicar-se a fazer o bem máximo ao mundo a partir da racionalidade extrema e lógica contraintuitiva para guiar decisões, havia, em grande parte e graças à influência de Bostrom, identificado a ameaça existencial de uma IA descontrolada como uma das principais áreas de atuação para seus seguidores. Dois anos antes, Holden Karnofsky, marido de Daniela, havia fundado a Open Philanthropy, uma organização sem fins lucrativos, para doar dinheiro, em parte, com base nos princípios do AE. A Open Phil, como era chamada, rapidamente se tornaria a principal financiadora de pesquisas sobre segurança da IA relacionadas a riscos catastróficos e existenciais. (Em novembro de 2024, havia[27] concedido mais de trezentos subsídios relacionados à segurança da IA, totalizando 440 milhões de dólares.)

Entretanto, tal vertente existencial da segurança da IA, baseada em experimentos filosóficos, logo viria a ser criticada à medida que a comunidade de pesquisa em inteligência artificial despertava para os danos reais, menos apocalípticos e mais imediatos, causados pela tecnologia. Por volta da mesma época em que Amodei[28] publicou seu artigo, a ProPublica divulgou uma investigação inovadora chamada *Machine Bias* [Predisposição das máquinas], revelando que algoritmos estavam sendo utilizados em todo o sistema de justiça criminal dos Estados Unidos em tentativas equivocadas de prever futuros criminosos, e que classificavam pessoas negras como de maior risco do que brancos que tinham registros criminais mais extensos. A matéria, somada a um ceticismo crescente em relação às big techs após 2016, sobre os danos das redes sociais, gerou uma nova onda de pesquisas voltadas a compreender os impactos sociais nocivos da inteligência artificial.

Deborah Raji, pesquisadora de responsabilidade em IA[29] na Universidade da Califórnia, Berkeley, viria a defender a reavaliação do foco excessivo das pesquisas de segurança da IA na teoria da inteligência artificial descontrolada e seus possíveis riscos existenciais, em detrimento e desfavorecimento de outros problemas reais, baseados em evidências, sendo coautora, em 2020, de um artigo em resposta ao de Amodei. Ela argumentou que sistemas de IA realmente "seguros" não poderiam ser criados isolando os comportamentos dos próprios sistemas técnicos sem inseri-los no contexto completo de seus impactos sobre aquilo que Amodei havia apontado — privacidade, equidade e economia. Enquanto ele levantava a ideia de que a IA poderia causar "efeitos colaterais negativos" ao perseguir implacavelmente um

objetivo, usando um exemplo similar ao do experimento mental do clipe de papel, no qual um robô de limpeza derruba um vaso ou danifica paredes no caminho em nome da limpeza, Raji apontou que isso já estava acontecendo. Na busca acirrada por produtos comerciais e IAG, a indústria da inteligência artificial havia gerado efeitos colaterais negativos de larga escala, incluindo a ampla violação da privacidade para treinar reconhecimento facial e os crescentes custos ambientais dos data centers necessários para sustentar o desenvolvimento da tecnologia.

"Não são apenas as *ações* de um agente da IA que podem produzir efeitos colaterais", escreveram ela e sua coautora. "Na vida real, escolhas básicas de design envolvidas na criação e no processo de implementação de modelos também têm consequências que vão muito além do impacto que a decisão de um único modelo pode ter. Na realidade, para que sistemas de IA ao menos sejam criados, há muito frequentemente um custo humano oculto."

Dentro da OpenAI, diversos pesquisadores, alguns deles entre o pequeno grupo de mulheres negras da empresa, pressionavam os executivos para expandir a definição de "segurança da IA" e incluir pesquisas em áreas como os impactos discriminatórios de modelos de aprendizagem profunda, ou deep learning. Os executivos mostraram-se indiferentes. "Esse não é nosso papel", disse um deles.

Em maio de 2016, Amodei, ainda no Google, passou pelo escritório da OpenAI para ver como as coisas estavam indo. A empresa acabara de se mudar do apartamento de Brockman para um espaço acima de uma fábrica de chocolate no Mission District, o bairro mais antigo de São Francisco e reduto latino da cidade. Os pesquisadores circulavam de meias pelo local.

"Há de vinte a trinta[30] pessoas na área, incluindo Nick Bostrom e o artigo da Wikipédia dizendo que o objetivo da OpenAI é desenvolver uma IA amigável e depois liberar seu código-fonte para o mundo", disse Amodei a Altman e Brockman, segundo relato na *New Yorker*.

"Não planejamos liberar todo nosso código-fonte", respondeu Altman. "Mas, por favor, não tentem corrigir isso. Normalmente, só piora a situação."

"Mas qual *é* o objetivo?", perguntou Amodei.

"Nosso objetivo no momento… é fazer o melhor possível", respondeu Brockman. "É um pouco vago."

Amodei se juntou dois meses depois para liderar a pesquisa em segurança da IA. Mais tarde, a Open Phil doaria[31] 30 milhões à OpenAI para garantir um assento no conselho por três anos para Holden Karnofsky. Em 2018, a convite de Brockman, Daniela, que havia sido a primeira recrutadora da Stripe, também iria para a empresa a fim de estruturar a equipe como gerente de engenharia e vice-presidente de recursos humanos. "Temos uma história longa[32] e fofa de como nos conhecemos", brincaria comigo sobre ela e Brockman um ano depois. "Isso mesmo", responderia Brockman, rindo. "Quando criamos a OpenAI, e comecei o recrutamento inicial aqui, pensei: 'Como eu gostaria de ter a Daniela na equipe.'"

No final de 2020, os irmãos Amodei ficariam tão perturbados com o que viam como o afastamento de Altman e da OpenAI de sua premissa original que decidiram criar a Anthropic, outro laboratório de IA, levando consigo funcionários importantíssimos e dando início a uma rivalidade que desempenharia um papel central no lançamento frenético do ChatGPT. Karnofsky deixaria o conselho da OpenAI após o cumprimento de seu mandato e devido ao novo conflito de interesses. Na lista de candidatos indicados para sua substituição, incluiria uma de suas ex-funcionárias: Helen Toner.

O problema era que a OpenAI não tinha ideia do que estava fazendo. Um ano depois de sua fundação, havia atraído, implorado e tomado emprestado o suficiente para montar uma equipe estelar na disputa agressiva por talentos dentro da indústria, mantendo o entusiasmo interno apenas pela quantidade de pessoas brilhantes. Ainda assim, a empresa lutava para encontrar uma estratégia coerente, e o ímpeto e o brilho começavam a se desgastar.

A lista de projetos se espalhava em todas as direções, refletindo de forma caótica o próprio campo. A OpenAI estava usando robôs, videogames e mundos virtuais simulados para treinar agentes — todas maneiras de tentar alcançar capacidades mais avançadas de inteligência artificial. Pouca coisa de fato funcionava, e o que funcionava mais parecia derivado de algo que alguém já havia criado. Fosse o que fosse a IAG, não era aquilo. “Os maiores projetos que eles tinham não pareciam fazer nada realmente inovador”, disse Nikhil Mishra, um pesquisador de IA que estagiou na OpenAI em 2017.

As habilidades de liderança de Brockman e Sutskever também estavam sendo testadas ao limite. Enquanto Brockman passava a maior parte dos dias programando, Sutskever circulava pelo escritório perguntando diversas vezes a cada pesquisador: “Qual é a sua próxima grande ideia?”, o que criava um ambiente sem rumo e de alto estresse. Não existia uma estrutura real de gestão nem um conjunto claro de prioridades. Às vezes, funcionários eram demitidos nos fins de semana, e o restante da equipe apenas tomaria conhecimento na segunda-feira seguinte, ao perceber as ausências. Além disso, o laboratório estava queimando dinheiro, a maior parte para manter os altos salários da equipe que havia sido montada. Em 2016, a OpenAI gastou[33] mais de 7 milhões de dólares dos 11 milhões em despesas apenas com remuneração e benefícios.

Musk estava perdendo a paciência. Não ajudava o fato de a DeepMind estar repentinamente sendo adulada no mundo inteiro. Em março de 2016, o programa AlphaGo derrotou Lee Sedol, um dos melhores jogadores humanos do mundo no antigo jogo chinês Go. (“A DeepMind está me causando um estresse mental extremo”, escreveu Musk à liderança da OpenAI pouco antes da série de cinco partidas. “Se eles vencerem, será uma péssima notícia, com essa filosofia que eles têm de uma mente apenas para governar o mundo.”) As partidas foram transmitidas ao vivo da Coreia do Sul para mais de 200

milhões de espectadores. Um ano depois, a Netflix lançaria um documentário de grande sucesso sobre a trajetória da empresa.

Musk aparecia de tempos em tempos no escritório para cobrar mais progresso, às vezes estabelecendo prazos totalmente irreais, traço característico de sua filosofia de gestão. Muitos funcionários se irritavam com essas expectativas, acreditando que não faziam sentido diante da natureza sinuosa e imprevisível da pesquisa. Durante uma reunião geral, Wojciech Zaremba, líder de robótica que fazia parte do grupo fundador, apresentou seus planos para os tipos de avanço em robótica que queria que sua equipe buscasse. Musk fez apenas uma pergunta: "Quando? Quando você vai fazer essas coisas?"

"Não sei", respondeu Zaremba.

Musk retrucou: "Bem, então você não tem um plano de verdade."

Foi assim que, em março de 2017,[34] Brockman e Sutskever começaram de fato a desenvolver um roteiro de pesquisa mais focado. A pergunta central era: o que realmente seria necessário para que a OpenAI alcançasse a IAG — e fosse a primeira a realizar a façanha?

Intuitivamente, Sutskever acreditava que tudo dependeria de uma dimensão acima de todas as outras: a quantidade de "*compute*", um termo técnico para recursos computacionais, de que a OpenAI precisaria para alcançar avanços nas capacidades da IA. A competição da ImageNet e os subsequentes avanços dos quais ele havia participado envolveram um aumento substancial na quantidade de *compute* utilizada para treinar um modelo de inteligência artificial. Tais avanços também envolveram outras coisas: um número significativo de mais dados e algoritmos mais sofisticados. No entanto, para Sutskever, o *compute* era rei. Se fosse possível ampliá-lo o bastante para treinar um modelo de IA na escala do cérebro humano, ele acreditava que algo radical certamente aconteceria: a IAG.

A quantidade de *compute* é baseada em três fatores: o poder de processamento de um chip individual, ou quantos cálculos consegue realizar por segundo; o número total de chips disponíveis; e por quanto tempo eles permanecem ligados para executar seus cálculos. O primeiro fator é determinado pela indústria de fabricação de chips, que, há décadas, dobra a capacidade de um chip a cada dois anos por meio de pesquisa e desenvolvimento intensivos. Essa taxa de progresso é conhecida como Lei de Moore, baseada em uma previsão que Gordon Moore, lendário cofundador da Intel, fez pela primeira vez nos anos 1960 — e que foi revisada uma década mais tarde — sobre a velocidade com que sua indústria poderia inovar. A Lei de Moore se tornou uma profecia autorrealizável, a meta de quão rápido as fabricantes de chips acreditavam *precisar* inovar para acompanhar a concorrência e permanecer relevantes.

Brockman e Sutskever realizaram um cálculo simples: com base no ritmo da Lei de Moore, quanto tempo levaria para atingir o nível de *compute* de que a OpenAI precisaria para uma IA em escala cerebral? A resposta foi deprimente: levaria tempo demais.

Na mesma época, Amodei[35] e Danny Hernandez, outro pesquisador, começaram a examinar a mesma ideia de uma perspectiva diferente. Em um gráfico simples, com o tempo no eixo X, eles inseriram a quantidade de *compute* usada em cada grande avanço em pesquisa de inteligência artificial desde 2012, começando com a descoberta de Sutskever na pós-graduação, o ponto de partida da revolução da IA. Descobriram que o uso de *compute* estava, na verdade, crescendo mais rápido do que a Lei de Moore. Muito mais rápido. Nos seis anos anteriores,[36] havia dobrado a cada *3,4 meses*, ou, em outras palavras, aumentado *30 milhões por cento*.

Brockman passou a chamar essa nova curva de duplicação de Lei da OpenAI. A liderança acreditava que, para atingir o objetivo final, a empresa não somente precisava de quantidades exorbitantes de *compute*, mas também escalar esse *compute* em um ritmo que, no mínimo, acompanhasse a nova lei. As fabricantes de chips haviam imposto a Lei de Moore a suas empresas com fervor existencial; agora, a liderança via a Lei da OpenAI sob o mesmo prisma.

Se não podiam esperar pela Lei de Moore, precisariam aumentar o *compute* de outro jeito: necessitariam de muito mais chips.

Os tipos de chips dos quais a OpenAI precisava eram caros. Conhecidos como unidades de processamento gráficos, ou GPUs, haviam sido originalmente projetados para renderizar em alta velocidade gráficos em computadores, dando, por exemplo, aos videogames um acabamento mais vivo e de baixa latência. No entanto, o mesmo formato se mostrava excelente para treinar os modelos de IA que a OpenAI queria desenvolver, já que compartilhava com a renderização gráfica uma exigência comum: a necessidade de processar quantidades enormes de números em paralelo.

A vasta maioria da indústria comprava essas GPUs de apenas uma empresa: a Nvidia, fabricante de chips sediada em Santa Clara, na Califórnia. A empresa não apenas produzia as melhores GPUs do mundo, como também havia desenvolvido uma plataforma de software complementar chamada CUDA, sigla para *Compute Unified Device Architecture*, que exercia um controle poderoso sobre os desenvolvedores de IA.

Em 2017, um servidor personalizado da Nvidia com oito de suas melhores GPUs custava 150 mil dólares — um preço que aumentaria consideravelmente com a inflação, chegando a quase 195 mil dólares em 2023. Nos anos seguintes, a Lei da OpenAI projetava que a empresa precisaria de milhares, se não de dezenas de milhares, de GPUs para treinar apenas um único modelo. O custo de eletricidade para alimentar esse treinamento também seria exorbitante. A OpenAI precisava de mais dinheiro — não apenas de 1 bilhão de dólares, mas de bilhões para se sustentar nos anos seguintes.

A constatação faria a empresa perder sua base financeira. Para Brockman e Sutskever, isso colocava em xeque a própria premissa da estrutura da OpenAI. Como uma organização sem fins lucrativos poderia levantar tanto dinheiro

anualmente para acompanhar o ritmo necessário a fim de permanecer na liderança? Por um breve período, eles chegaram a considerar uma fusão[37] com uma startup de chips, mas, no verão de 2017, começaram discussões sérias com Altman e Musk sobre a necessidade de transformar a OpenAI em uma empresa com fins lucrativos. Essa era a melhor opção para atrair investidores com a perspectiva de obter retorno financeiro. Após diversas semanas de negociações, as deliberações terminaram bruscamente, sem resolução. Se a OpenAI se transformasse em uma empresa com fins lucrativos, Altman, que estava considerando concorrer ao governo da Califórnia e não era muito bem aceito nos grupos focais, queria ser o diretor-executivo. Musk queria o mesmo[38] posto; exigia controle total do laboratório e posse majoritária das ações.

No fogo cruzado, Sutskever e Brockman quase optaram por Musk, já que ambos preferiam sua liderança. No entanto, Altman fez um apelo direto a Brockman, evocando a relação interpessoal e levantando preocupações sobre a imprevisibilidade de Musk. Ele enfrentava muitas pressões externas e era propenso a comportamentos erráticos e instáveis. Caso a OpenAI fosse bem-sucedida, não seria perigoso dar a Musk controle total da IAG? Convencido, Brockman apelou para Sutskever, que continuava em dúvida. Em setembro de 2017, Sutskever enviou um e-mail a Musk e Altman, em seu nome e no de Brockman, em uma última tentativa de resolver a situação, dizendo:

"**Elon:** Nós *realmente* queremos trabalhar com você. Acreditamos que, se unirmos forças, nossas chances de sucesso na missão serão maiores." No entanto, o desejo de Musk por controle total parecia antiético em relação ao espírito original da OpenAI. "Você está preocupado que Demis possa criar uma ditadura da IAG. Temos a mesma apreensão. Por isso, é uma má ideia criar uma estrutura em que você poderia se tornar um ditador, caso fosse de sua vontade."

"**Sam:** Quando Greg e eu ficamos sem saída, você sempre teve uma resposta que acabou sendo profunda e correta", prosseguiu Sutskever. Dito isso, os comportamentos de Altman muitas vezes deixavam os dois confusos sobre suas verdadeiras crenças e intenções. "Nós não entendemos por que o título de CEO é tão importante para você. Suas razões declaradas mudaram, e é difícil compreender o que de fato está motivando isso. A IAG é *realmente* sua principal motivação? Como isso se conecta a seus objetivos políticos? Como foi que sua linha de pensamento mudou ao longo do tempo?"

"Há bagagem suficiente aqui para considerarmos muito importante nos reunirmos para conversarmos abertamente", concluiu o e-mail. "Se todos formos honestos e resolvermos as questões pendentes, a empresa que criaremos terá chances muito maiores de resistir às forças muito intensas que terá de enfrentar em seu caminho."

Em dez minutos, Musk respondeu: "Pessoal, já deu. Esta foi a gota d'água", escreveu. Se Sutskever e Brockman ainda quisessem seguir o caminho de uma empresa com fins lucrativos, teriam que fazer isso por conta própria;

caso contrário, a OpenAI continuaria como uma organização sem fins lucrativos. “Não financiarei mais a OpenAI até que vocês assumam um compromisso firme de permanecer, ou então serei apenas um tolo que está, basicamente, financiando uma startup sem nada em troca”, disse Musk.

Cinquenta minutos depois, ele enviou outra mensagem. “Para deixar claro, isto não é um ultimato para aceitar o que já foi discutido. O assunto não está mais em pauta.”

Na manhã seguinte, Altman apareceu na mesma troca de e-mails: “Continuo animado com a estrutura sem fins lucrativos!” Ele também enviou garantias adicionais a Musk por meio de Shivon Zilis, uma das assessoras de confiança de Musk, que trabalhava na Tesla e na Neuralink, sua empresa de interfaces neurotecnológicas implantáveis. “É ótimo manter a OpenAI como uma organização sem fins lucrativos e continuar apoiando”, escreveu Zilis a Musk, com anotações do que Altman lhe passara. “Ele admitiu que perdeu muita confiança em Greg e Ilya durante esse processo. Sentiu que a comunicação de ambos foi inconsistente e, às vezes, infantil.” Altman também havia se incomodado com o quanto os dois compartilhavam com o restante da OpenAI durante as negociações. “Acho que isso distraía a equipe”, disse Zilis.

Entretanto, a realidade era que manter a OpenAI como uma organização sem fins lucrativos não resolveria o problema financeiro. Enquanto Brockman e Sutskever continuavam[39] a se reunir com potenciais investidores filantrópicos, tinham dificuldade em se aproximar do tipo de capital que acreditavam ser necessário para sustentar a empresa. A oscilação caprichosa de Musk em seu compromisso de financiamento também ameaçava lançá-la em uma crise. Nos bastidores, Altman começou a buscar alternativas de financiamento e a tentar reduzir a dependência da OpenAI em relação a Musk. Ligou para Reid Hoffman,[40] que se ofereceu para intervir e cobrir os salários dos funcionários e os custos operacionais. Altman considerou lançar[41] uma nova criptomoeda. Investigou uma série de diferentes estruturas corporativas, incluindo uma corporação de benefício público, modelo que Musk defendia e que permitiria à OpenAI se tornar uma empresa com fins lucrativos, mas ainda legalmente vinculada a sua missão.

A urgência se acentuava devido a uma preocupação constante: a possibilidade de a OpenAI perder seus melhores pesquisadores a qualquer instante. Antes, com o forte apoio financeiro de Musk,[42] a empresa havia aumentado agressivamente os salários dentro do formato sem fins lucrativos a fim de evitar contraofertas. Porém, então, a guerra por talentos apenas se intensificara, e o próprio Musk havia, em junho de 2017, tirado da OpenAI Andrej Karpathy, um de seus cientistas fundadores mais importantes, para dirigir a divisão de IA da Tesla. Em termos de remuneração, a OpenAI tinha uma grande desvantagem: não podia oferecer participação acionária, algo que muitos profissionais de tecnologia do Vale do Silício consideravam essencial para arcar com o alto custo de vida da região.

Musk logo chegou à própria conclusão no que dizia respeito à solução do problema financeiro da OpenAI. Em janeiro de 2018, Andrej Karpathy enviou-lhe um e-mail com novos dados mostrando o quanto o Google dominava as principais publicações de pesquisa em inteligência artificial. "Trabalhar na vanguarda da IA é, infelizmente, oneroso", escreveu. "Parece-me que, hoje, a OpenAI está queimando dinheiro e que o modelo de financiamento não consegue atingir a escala necessária para competir seriamente com o Google (uma empresa de 800 bilhões de dólares). "Transformar a OpenAI em uma empresa com fins lucrativos poderia ajudar a levantar capital, mas demandaria que o laboratório desenvolvesse um produto de IA do zero, uma distração significativa de sua pesquisa principal. "A opção mais promissora que consigo imaginar, como mencionei antes, seria a OpenAI se associar à Tesla como sua vaca leiteira", continuou Karpathy. A Tesla já havia feito boa parte do trabalho pesado para desenvolver um produto de inteligência artificial — o Autopilot, sua função de piloto automático. Caso a OpenAI pudesse ajudar a acelerar os esforços da Tesla para transformar o Autopilot em uma solução completa de direção autônoma, isso por si só poderia aumentar a receita da Tesla o bastante para arcar com os custos altíssimos da computação da OpenAI.

Musk encaminhou o e-mail de Karpathy para Brockman e Sutskever. "Andrej está coberto de razão", escreveu. "A Tesla é o único caminho que poderia, ao menos, ter alguma chance de rivalizar com o Google. Ainda assim, a probabilidade de ser um contrapeso de verdade é pequena. Só não é nula."

Entretanto, àquela altura, Altman havia abandonado seus planos políticos e conseguido convencer Brockman, e por intermédio dele, Sutskever, de que seria o melhor líder para a organização. Com tal decisão, Musk não quis mais ser publicamente associado à OpenAI. "Não ficarei em uma situação na qual a percepção sobre minha influência e meu tempo não corresponda à realidade", já havia escrito. Algumas semanas depois, Musk renunciou ao cargo de copresidente da OpenAI, e Altman se tornou presidente da organização sem fins lucrativos.

Para o público, a OpenAI apresentou a saída de Musk como resultado de um conflito de interesses, mantendo silêncio sobre sua nova realidade financeira: do bilhão prometido, a empresa recebeu, no fim das contas, apenas cerca de 130 milhões de dólares, dos quais menos de 45 milhões vieram de Musk. O futuro da OpenAI passou a depender exclusivamente da habilidade de Altman de captar recursos para compensar tais perdas e continuar atendendo à demanda crescente de capital.

Musk anunciou pessoalmente sua decisão durante uma reunião geral da OpenAI. Para muitos funcionários, que não estavam cientes do drama que acontecia nos bastidores da liderança, sua saída trouxe alívio, mas também uma grande incerteza quanto ao futuro da organização. Até então, Musk tinha sido um dos principais responsáveis pela visibilidade pública do laboratório. Durante a reunião, ele não se conteve: disse que, em primeiro lugar, era

primordial desenvolver uma IAG segura e que estava claro que a OpenAI falharia nesse objetivo enquanto permanecesse uma organização sem fins lucrativos. Afirmou que passaria a buscar o mesmo propósito na Tesla, que, segundo ele, tinha chances muito maiores de sucesso graças aos recursos abundantes de uma empresa bem financiada.

Um estagiário questionou as intenções de Musk. Seria aquela a melhor solução de fato? Ele havia realmente esgotado todas as alternativas? Avançar com um concorrente da OpenAI dentro da Tesla parecia apenas fomentar uma corrida comercial e poderia colocar em risco o desenvolvimento seguro da IAG. "Isso não é voltar atrás no que você mesmo disse não querer fazer?", perguntou o estagiário.

Musk explodiu. "Você é um idiota! Eu já pensei nisso mais do que você pode imaginar. Tentei de tudo. Você não faz ideia de quanto tempo passei pensando neste assunto", disse. "Estou realmente assustado com tudo isso."

Mais tarde, o estagiário foi homenageado[43] por sua coragem com um troféu de "idiota".

No dia seguinte ao Natal daquele ano, Musk escreveu novamente a Altman, Brockman e Sutskever:

> ASSUNTO: Sinto que devo repetir.
>
> A avaliação que fiz da probabilidade de a OpenAI ser relevante em relação à DeepMind/Google, sem uma drástica mudança de execução e recursos, é de 0%. Não de 1%. Gostaria que fosse diferente.
> Arrecadar centenas e centenas de milhões não será suficiente.
> Isso exige bilhões por ano imediatamente, ou podem esquecer.

Altman precisava levantar fundos, o mais rápido possível.

A OpenAI intensificou sua estratégia de publicidade, concentrando-se em projetos de demonstração capazes de exibir suas capacidades para um público leigo. Um deles recebeu destaque: a tentativa de criação de um agente de IA capaz de derrotar os melhores jogadores humanos do mundo no complexo jogo de estratégia *Dota 2*. A empresa já havia desenvolvido um agente que podia vencer o melhor jogador humano em partidas contra um. A meta, então, seria formar uma equipe de cinco agentes para enfrentar o melhor time humano de cinco jogadores do mundo.

Conscientemente ou não, era uma reprise da estratégia da DeepMind. O campeonato mundial de *Dota 2* seria transmitido ao vivo, oferecendo uma vitrine perfeita para o trabalho da OpenAI, em termos simples e dramáticos de vitória ou derrota. Enquanto isso, a DeepMind já trabalhava em um projeto parecido com o jogo *StarCraft II*, o que criava um ponto de comparação natural, ainda que arbitrário, para possíveis investidores da OpenAI. Além disso, o projeto de *Dota 2* exigia um enorme poder computacional, servindo

também como um teste prático da estratégia de escalonamento da empresa. Brockman, que liderou a fase inicial do projeto, ampliou sua equipe e começou o trabalho.

Tudo o que faltava então era um documentário.

A tarefa ficou a cargo de um integrante da equipe de robótica da OpenAI. Ele comprou equipamentos de filmagem caros e começou a seguir o time de *Dota 2* dentro do escritório. Escreveu o próprio roteiro e montou um corte bruto de três horas, uma verdadeira saga. Contudo, apesar de todo o esforço, quem assistiu à versão não editada concordou que era terrível. Sendo assim, profissionais foram contratados, e Brockman[44] passou a financiá-los, em parte com o próprio dinheiro.

Enquanto isso, Altman detalhava o plano para arrecadar dinheiro. Após considerar uma variedade de estruturas com fins lucrativos, chegou a uma proposta incomum para equilibrar a necessidade de capital com o compromisso contínuo com a missão da OpenAI. Embora as empresas de benefício público tivessem um mecanismo embutido para manter tal equilíbrio, elas também vinham com muitas outras regras. Em vez disso, Altman criaria uma sociedade limitada — ou LP, para *limited partnership* — com o intuito de atuar como o braço com fins lucrativos destinado a receber investimentos e comercializar as tecnologias da OpenAI. Esse braço estabeleceria um teto aos retornos dos investidores e seria controlado pela entidade sem fins lucrativos da empresa. A vantagem era que o acordo operacional de uma LP podia ser redigido de acordo com os desejos do criador. A OpenAI poderia especificar que a missão teria prioridade sobre os investidores. As sociedades limitadas também determinavam o poder que os acionistas poderiam exercer, impedindo que alcançassem controle majoritário.

Altman apresentou a proposta aos funcionários com cuidado: o compromisso inicial da OpenAI de evitar motivações lucrativas havia sido feito com o objetivo de impedir que o laboratório comprometesse sua missão. Porém, considerando que seu sucesso demandava um capital que a entidade sem fins lucrativos não podia captar, manter a estrutura original representava então um risco maior para a própria missão. No fim, a maioria concordou, ainda que com a relutância de alguns, que a LP era o melhor caminho a seguir.

Em abril de 2018, a OpenAI[45] publicou uma carta de princípios para abrir caminho para a transição. Sem revelar publicamente nenhum detalhe sobre a mudança que estava por vir, o documento reiterava o propósito do laboratório, mas de um modo reformulado: "A missão da OpenAI é garantir que a inteligência artificial geral (IAG)… beneficie toda a humanidade." O documento acrescentava que uma missão como essa exigiria que a empresa estivesse na "dianteira das capacidades da IA" e que isso demandaria "recursos substanciais"; também poderia significar um recuo no compromisso de divulgar as pesquisas de seu laboratório por conta de "preocupações com segurança e proteção". Pela primeira vez, a OpenAI também definiu formalmente o que entendia por IAG: "Sistemas extraordinariamente

autônomos que superam os seres humanos na maioria dos trabalhos de valor econômico."

Naquele verão, enquanto a equipe do *Dota*[46] começava a vencer partidas amadoras e a divulgar seus resultados na mídia de tecnologia ("IA da OpenAI para *Dota 2* atropela equipe campeã mundial de e-sports com vitórias consecutivas", exaltava uma manchete), Altman se encontrou com Satya Nadella, CEO da Microsoft, na conferência da Allen & Company em Sun Valley, Idaho. O evento anual, conhecido como o "acampamento de verão para bilionários", havia sido palco de diversos grandes acordos corporativos. Altman estava pronto para fechar o seu.

Ele apresentou a Nadella uma proposta de investimento na OpenAI, suficiente para despertar o interesse do executivo-chefe. No entanto, Nadella questionou se deveria investir em uma organização externa, quando a própria empresa já tinha uma divisão de pesquisa em IA de longa data dentro da Microsoft Research. Ao retornar para a Microsoft, a questão foi levada a seus conselheiros seniores.

"A Microsoft Research e a OpenAI[47] são ambas organizações que avançam na fronteira da tecnologia", ponderou Xuedong Huang, então diretor de tecnologia da Azure AI. Por que não investir nas duas?

Em menos de seis meses, a OpenAI e a Microsoft já discutiam seriamente um acordo. Altman cuidou da parte legal, apressando a criação da sociedade limitada e nomeando-se CEO. Internamente, o projeto tinha o codinome "Oregon Trail". Para manter o acordo afastado de olhares curiosos,[48] a entidade com fins lucrativos também foi incorporada sob o pseudônimo SummerSafe LP. O nome fazia referência a um episódio do desenho animado *Rick e Morty*, em que os personagens principais, o cientista louco Rick e seu neto Morty, deixam Summer, a irmã mais velha do menino, em outro universo e instruem seu carro a "mantê-la segura" ("*keep Summer safe*"). O carro leva a missão a sério, recorrendo a mecanismos de defesa extremos e prejudiciais, incluindo assassinato, paralisia e tortura de pessoas que se aproximam do veículo. Era uma alusão às potenciais armadilhas da IA.

No início de 2019, a liderança sênior da Microsoft começou a passar pelo escritório da OpenAI. O primeiro foi Kevin Scott, animado diretor de tecnologia da multinacional, que acompanhava a OpenAI e se mostrava particularmente afeiçoado à startup; em seguida, Craig Mundie, conselheiro sênior de Nadella e membro da liderança da Microsoft por mais de vinte anos, inclusive como diretor de pesquisa e estratégia. Bill Gates também apareceu, reservado e comedido como sempre, observando uma série de demonstrações. A maior parte dos funcionários ficou sem saber do envolvimento da Microsoft. Altman pediu à pequena equipe que trabalhava no acordo que não divulgasse informações sobre o possível investimento.

Na mesma época,[49] Altman começou a enfrentar problemas na YC. Depois de cinco anos à frente da organização, a frustração com ele havia

atingido níveis críticos por uma questão extremamente similar à que surgira na Loopt: a aparente priorização dos próprios projetos e ambições acima dos da organização — às vezes, até mesmo em prejuízo dela. O tempo que ele dedicava às negociações da OpenAI em vez de aconselhar as startups da YC não ajudava. Alguns o viam como alguém que colhia benefícios pessoais significativos, obtendo retornos exorbitantes ao investir em empresas da YC com a Hydrazine, seu fundo pessoal, enquanto realizava trabalho limitado. Ao tomar conhecimento de sua ausência, a preocupada Jessica Livingston incitou Altman a renunciar à presidência da YC, segundo o *Washington Post*. Ele concordou e, no início de 2019, Paul Graham viajou do Reino Unido, onde havia se aposentado, para São Francisco, com o intuito de finalizar a decisão.

Altman tentou suavizar publicamente a mudança. Em 8 de março de 2019, no mesmo dia em que recebeu o senador Schumer, publicou um post no blog da YC anunciando que faria a transição da presidência da empresa para a presidência do conselho de administração, a fim de dedicar mais tempo à OpenAI. Dias depois, em 11 de março, Brockman e Sutskever apresentaram publicamente a OpenAI LP, a sociedade limitada da organização, e Altman revelou seu papel como CEO. O timing foi estratégico. A mídia reportou extensivamente a mudança como um passo bem coreografado em sua carreira e seu novo papel como membro do conselho da YC. Contudo, na verdade, ele não detinha o título; havia proposto a ideia[50] a seus pares da YC, mas a divulgou como se fosse uma conclusão já acertada, sem a decisão deles, reportou o *Wall Street Journal*. O post no blog foi posteriormente editado para remover qualquer menção a Altman.

Na OpenAI, seu novo título apenas formalizou o papel que ele já desempenhava desde a saída de Musk. Quando assumiu as rédeas, muitos funcionários se sentiram aliviados. Sua postura calma e controlada era um contraste bem-vindo à intensidade e aos humores imprevisíveis de Musk. Altman também ajudou a amenizar as crescentes reclamações sobre a gestão de Brockman e Sutskever. Ele levou um coach para executivos e ofereceu treinamento aos gerentes. Além disso, instalou mais líderes seniores, incluindo Brad Lightcap, investidor da YC, como diretor financeiro; promoveu Bob McGrew, que antes liderava engenharia e gerenciamento de produtos na Palantir, fundada por Thiel, da equipe de robótica para VP de pesquisa; e contratou Mira Murati, líder de produção e engenharia na Leap Motion, startup de realidade virtual, e no Model X da Tesla, para supervisionar a estratégia de hardware e uma linha central de pesquisa.

Com a formação da OpenAI LP, a maioria dos funcionários deixou a organização sem fins lucrativos e assinou novos contratos, com participação acionária, com o braço com fins lucrativos. (As exceções incluíam funcionários internacionais com vistos vinculados ao setor original.) Uma estrutura de faixas salariais[51] vinculava a remuneração não apenas à "expertise em engenharia" e à "direção de pesquisa", mas também à conformidade com o estatuto. Funcionários de nível três precisavam "entender e internalizar o

estatuto da OpenAI". Os de nível cinco precisavam "garantir que todos os projetos em que você e sua equipe trabalhem estejam consistentes com o estatuto". Os de nível sete eram "responsáveis por defender e melhorar o estatuto, e por responsabilizar outros na organização a fazer o mesmo". Executivos também elaboraram[52] um FAQ para acalmar tensões remanescentes. "Posso confiar na OpenAI?", dizia uma das perguntas. A resposta começava com "Sim".

No mundo mais amplo da tecnologia, a transição da OpenAI desencadeou uma onda de acusações de que o laboratório estava retrocedendo em sua promessa original. Os termos iniciais da sociedade limitada estabeleciam que a primeira rodada de investidores teria seus retornos limitados a cem vezes o valor investido. A OpenAI chamou a estrutura criada de "empresa de lucro limitado". Em uma publicação no *Hacker News*, um site popular de agregação de notícias administrado pela YC, um usuário questionou o quanto esse limite era de fato significativo. "Então alguém que investe[53] 10 milhões de dólares tem seu investimento 'limitado' a um bilhão. Rá, rá. Basicamente ilimitado, a menos que a empresa cresça até o valor de mercado de uma FAANG", escreveu, usando o acrônimo para Facebook, Apple, Amazon, Netflix e Google.

Brockman respondeu como "gdb", seu nome de usuário: "Acreditamos que, se realmente criarmos uma IAG, geraremos uma ordem de magnitude de valor muito maior do que qualquer empresa existente."

Outro usuário respondeu em seguida. "Os primeiros investidores do Google tiveram um retorno de cerca de vinte vezes sobre seu capital. A empresa atualmente está avaliada em 750 bilhões de dólares. Sua aposta é que vocês terão uma estrutura corporativa que gerará *ordens de magnitude* maiores do que o Google… mas não quererem 'concentrar poder indevidamente'?", escreveu, citando o estatuto. "O que exatamente é poder, senão a concentração de recursos?"

Os investimentos iniciais começaram a fluir[54] para a LP, incluindo mais de 60 milhões de dólares transferidos da OpenAI sem fins lucrativos, 10 milhões da YC e 50 milhões cada da Khosla Ventures e da fundação filantrópica de Hoffman. A princípio, Hoffman estava relutante[55] em investir mais na OpenAI, já que a empresa não tinha produto nem plano de mercado, como ele contaria depois. No entanto, acabou concordando em coliderar a rodada depois de Altman lhe dizer que isso ajudaria a legitimar a seriedade da intenção da empresa de desenvolver um negócio lucrativo.

Enquanto isso, a Microsoft continuava em deliberação. Nadella, Scott e outros executivos da empresa já estavam de acordo com um investimento inicial. O único a resistir era Bill Gates.

Para Gates, o *Dota 2* não era nada empolgante. Ele também não se impressionava com robótica. A equipe da área havia criado a demonstração de uma mão robotizada que aprendera a resolver um cubo mágico por tentativa e

erro, o que havia recebido cobertura amplamente positiva. Gates não via utilidade nisso; queria um modelo de IA que pudesse ler livros, compreender conceitos científicos e responder a perguntas com base no material — um assistente para conduzir pesquisas.

A OpenAI tinha apenas um projeto que se aproximava do que ele queria: um grande modelo de linguagem chamado GPT-2, capaz de gerar trechos de textos que se assemelhavam bastante à escrita humana. Em fevereiro daquele ano, a OpenAI havia tomado a incomum medida de anunciar à imprensa que esse modelo, assim que estivesse um pouco mais avançado, poderia se tornar uma tecnologia extremamente perigosa. Governos autoritários ou organizações terroristas poderiam transformá-lo em arma para produzir desinformação em massa. Usuários poderiam inundar a internet com uma quantidade imensurável de conteúdo inútil, a ponto de dificultar encontrar informações de qualidade. A OpenAI disse que escolheria o caminho ético e reteria a versão completa do modelo, que tinha 1,5 bilhão de parâmetros, ou variáveis, uma medida aproximada do tamanho e da complexidade de um modelo. Em vez disso, para dar ao público apenas uma amostra das capacidades para as quais a sociedade precisava se preparar, publicaria apenas uma versão reduzida, com menos de um décimo do tamanho, capaz de gerar algumas frases por vez, mas propensa a conclusões ilógicas e repetições.

O GPT-2 ainda estava longe de compreender conceitos científicos, mas o modelo conseguia realizar certa sumarização básica de documentos e, de certo modo, responder a perguntas. Talvez, pensaram alguns pesquisadores da OpenAI, caso treinassem um modelo maior, com mais dados, e para executar tarefas que ao menos parecessem mais próximas ao que Gates queria, seriam capazes de convencê-lo a deixar o papel de crítico e torná-lo, no mínimo, neutro. Em abril de 2019, um pequeno grupo de pesquisadores foi até Seattle para fazer o que chamaram de "Demonstração a Gates", uma versão mais avançada do GPT-2. Por fim, ele de fato se convenceu o suficiente para que o acordo fosse fechado.

Em uma reunião geral subsequente, Altman anunciou a novidade, apresentando a Microsoft como o investidor e parceiro ideal. A gigante da tecnologia tinha o capital e o poder computacional necessário à OpenAI, e sua liderança estava profundamente alinhada com a missão de garantir uma IA geral benéfica. A OpenAI também havia feito compromissos bastante vagos sobre o que entregaria à Microsoft para fins de comercialização. O laboratório não precisou abrir mão de quase nada, disse Altman. Era um acordo muito bom.

Dentro da Microsoft, o investimento foi encarado de maneira prática. Se a OpenAI alcançaria ou não a IAG, não era realmente a questão. No entanto, a empresa estava claramente na dianteira, e investir cedo poderia, afinal, transformar a Microsoft em uma líder de inteligência artificial — tanto em software quanto em hardware —, em pé de igualdade com o Google. "O que é interessante[56] no que a OpenAI, a DeepMind e o Google Brain estão

fazendo é a escala de sua ambição", escreveu Scott a Nadella e Gates em meados de junho, referindo-se à divisão de pesquisa em IA do Google, "e como essa ambição está impulsionando tudo, desde o design de data centers até os chips de computação, redes, arquiteturas de sistemas distribuídos, otimizadores numéricos, compiladores, frameworks de programação e as abstrações de alto nível disponíveis aos desenvolvedores de modelos." Scott disse que a Microsoft estava incrivelmente atrasada em diversas frentes: tinha dificuldade em reproduzir os melhores modelos de linguagem do Google, e a Azure, sua plataforma de computação em nuvem, apresentava grandes lacunas em comparação com a infraestrutura equivalente da empresa. Levaria anos para a Microsoft alcançar o mesmo nível por conta própria. Ele estava "muito, muito preocupado".

Nadella respondeu no mesmo dia, retirando Gates da troca de e-mails e adicionando Amy Hood, diretora financeira da Microsoft. "Ótimo e-mail que explica por que quero que façamos isso…e por que garantiremos que nosso pessoal de infra coloque isso em ação", escreveu ele, usando a abreviação para *infraestrutura*.

Um mês depois, em 22 de julho de 2019, a Microsoft anunciou seu investimento de 1 bilhão de dólares. Pelos termos do acordo, os retornos seriam limitados a vinte vezes o valor investido.

CAPÍTULO 3

# CENTRO NERVOSO

Cheguei aos escritórios da OpenAI duas semanas depois, em 7 de agosto de 2019. Naquele tempo, o laboratório já havia se instalado em um prédio independente, não muito longe da antiga fábrica de chocolate, na esquina das ruas 18th e Folsom, em São Francisco. A fachada cinza exibia letras anunciando o prédio como um marco histórico: o edifício que um dia abrigara a Pioneer Tronque Factory. Três anos antes, Musk havia alugado o local por meio de uma de suas empresas, herdando o interior reformado de um espaço de coworking antes ocupado, em sua maioria, pela Stripe. A OpenAI mudou-se para lá com outro empreendimento do bilionário, a Neuralink, voltada para o desenvolvimento de interfaces cérebro-máquina.

Eu suava frio depois de cruzar o Mission District, passando por taquerias tradicionais que vinham sendo lentamente substituídas por novos cafés da moda e desviando da crescente população de pessoas em situação de rua. Uma rajada de ar frio me envolveu assim que entrei pela porta. Após a recepção, o hall de entrada se abria para um amplo lounge. A luz do sol atravessava as janelas, banhando o ambiente com tetos de vigas de madeira expostas e sofás acolhedores. À direita, uma cafeteria servia refeições aos funcionários; jogos de tabuleiro e livros se espalhavam pelas prateleiras que revestiam as paredes.

No Vale do Silício, o design dos escritórios é uma espécie de moeda, um símbolo de confiança no futuro financeiro da empresa e uma forma de conquistar uma pequena vantagem na disputa por talentos de alto nível. Em 2021, a OpenAI assumiria dois edifícios geminados a poucos quarteirões dali, investindo 10 milhões de dólares ao longo de dois anos para reformar mais de 30 mil metros quadrados.[1]

Enquanto os funcionários chamavam o primeiro escritório de Pioneer Building, o segundo, uma antiga fábrica de maionese, seria apelidado de Mayo. Altman supervisionaria pessoalmente a reforma do novo espaço, trocando a escada de estrutura metálica industrial do Pioneer por uma peça central ondulada de madeira e pedra; e as poltronas de couro, que podiam custar cerca de 2 mil dólares, por cadeiras de designers brasileiros avaliadas em mais de 10 mil dólares.

Ele também acrescentaria ao Mayo[2] um ambiente com prateleiras de madeira e um tapete persa, inspirado em uma combinação entre sua livraria parisiense favorita e uma sala de estudos da maior biblioteca[3] de sua *alma mater*, a Stanford. Altman desejava "um elemento aquático", contou ao designer de sua empresa, que chegou a propor uma magnífica cascata flutuante no centro do escritório, uma estrutura artificial que evocaria a harmonia entre homem e máquina. No fim, Altman seguiu outro caminho: instalou fontes de pedra borbulhantes cercadas por uma profusão de plantas, aninhadas ao redor dos sofás, pendendo do teto e descendo em cascata pelas paredes.[4]

Fora dos dois escritórios, o mesmo período de dois anos testemunharia a pandemia devastar ainda mais o que já se tornara o epicentro da gentrificação impulsionada pela indústria tecnológica. Negócios latinos de longa data fechariam as portas. Os crimes violentos aumentariam. Uma fileira de barracas e lixo acumulado surgiria a poucos passos do Pioneer Building, à medida que a falta de moradia atingisse níveis críticos.

Contudo, ali dentro, envolto em um brilho acolhedor, com revistas espalhadas sobre as mesas, era fácil acreditar em uma realidade mais suave. Uma funcionária me contaria mais tarde que aquilo era emblemático de sua passagem pela empresa: entrar lá era como adentrar um universo alternativo. Só depois de sair é que ela voltou à realidade.

Brockman, então com 31 anos, diretor de tecnologia da OpenAI e futuro presidente da empresa, desceu as escadas para me receber. Apertou minha mão com um sorriso hesitante.

"Nunca demos tanto acesso a ninguém", disse ele.

Naquele momento, poucas pessoas fora do restrito mundo da pesquisa em IA conheciam a OpenAI. No entanto, como repórter da *MIT Technology Review*, acompanhando de perto as fronteiras cada vez mais amplas da inteligência artificial, eu seguia atentamente seus passos.

Até aquele ano, a OpenAI era vista como uma espécie de agregada da pesquisa em IA. Tinha uma premissa ousada, a de que a IAG poderia ser alcançada em uma década, algo que a maioria dos especialistas fora do laboratório considerava impossível. Para grande parte das pessoas da área, a empresa dispunha de uma quantia obscena de recursos, mas pouca direção, e gastava somas elevadas promovendo o que outros pesquisadores viam como trabalhos sem originalidade. Para alguns, era também objeto de inveja. Como organização sem fins lucrativos, afirmava não ter interesse em buscar a comercialização. Era um raro playground intelectual sem amarras, um refúgio para ideias marginais.

No entanto, nos seis meses que antecederam minha visita, uma rápida sucessão de acontecimentos na OpenAI indicava uma grande mudança em sua trajetória. Primeiro, veio a decisão confusa de reter o GPT-2 e se vangloriar disso. Em seguida, o anúncio de que Altman, que havia deixado misteriosamente seu influente cargo na YC, assumiria a função de CEO da

OpenAI com a criação de sua nova estrutura de "lucro limitado". Eu já havia preparado minha visita quando a empresa revelou o acordo com a Microsoft, que concedia à gigante da tecnologia prioridade na comercialização das inovações da OpenAI e a obrigava a usar exclusivamente o Azure, a plataforma de computação em nuvem da própria Microsoft.

Cada novo anúncio gerava controvérsias, especulações intensas e atenção crescente, que começava a ultrapassar os limites da indústria de tecnologia. Enquanto meus colegas e eu acompanhávamos a evolução da empresa, era difícil compreender plenamente o alcance do que estava acontecendo. O que se tornava evidente, porém, era que a OpenAI começava a exercer uma influência significativa sobre a pesquisa em IA e sobre a maneira como os formuladores de políticas passavam a compreender a tecnologia. A decisão do laboratório de se tornar uma empresa parcialmente voltada ao lucro teria efeitos em cascata sobre suas esferas de influência na indústria e no governo.

Tarde da noite, por insistência de meu editor, enviei um e-mail a Jack Clark, diretor de políticas da OpenAI, com quem eu já havia falado: eu ficaria na cidade por duas semanas e aquele parecia o momento certo na história da OpenAI. Será que eles se interessariam por um perfil? Clark me encaminhou ao chefe da área de comunicação, que respondeu. A OpenAI estava, de fato, pronta para se reapresentar ao público. Eu teria três dias para entrevistar a liderança e me integrar à empresa.

Brockman e eu nos acomodamos em uma sala de reuniões envidraçada com Sutskever. Sentados lado a lado em uma longa mesa de conferência, cada um assumiu o papel esperado. Brockman, o programador e executor, inclinava-se levemente para a frente, um pouco tenso, ansioso por causar boa impressão; Sutskever, o pesquisador e filósofo, recostava-se na cadeira, relaxado e distante.

Abri o laptop e dei uma olhada nas minhas perguntas. A missão da OpenAI é garantir uma IAG benéfica, iniciei. Por que gastar bilhões de dólares nesse problema, e não em outro?

Brockman assentiu com entusiasmo. Estava acostumado a defender a posição da OpenAI.

"A razão pela qual nos preocupamos tanto com a IAG e achamos importante desenvolvê-la é porque acreditamos que ela pode ajudar a resolver problemas complexos que estão fora do alcance dos seres humanos", disse ele.

Deu dois exemplos que se tornaram quase dogmas entre os adeptos da IAG. Mudanças climáticas:

"É um problema supercomplexo. Como poderíamos resolvê-lo?"

E medicina:

"Veja como a saúde é uma questão política importante nos Estados Unidos atualmente. Como podemos realmente oferecer um tratamento melhor às pessoas a um custo menor?"

Sobre o último exemplo, ele começou a contar a história de um amigo que sofria de uma doença rara e recentemente havia passado por uma exaustiva

maratona com diferentes especialistas para descobrir o que havia de errado. A IAG, explicou, reuniria todas essas especialidades. Pessoas como seu amigo não precisariam mais desperdiçar tanta energia e frustração para obter uma resposta.

"Por que precisamos da IAG para fazer isso, em vez da IA?", perguntei.

Era uma distinção importante. O termo IAG, antes restrito a uma obscura seção do vocabulário tecnológico, começara apenas recentemente a ganhar uso mais amplo — em grande parte por causa da própria OpenAI. Conforme definido pela empresa, IAG se referia a um ápice teórico da pesquisa em inteligência artificial: um software com tanta sofisticação, agilidade e criatividade quanto a mente humana, capaz de igualar ou superar seu desempenho na maioria das tarefas com mais valor econômico.

A palavra-chave era *teórico*. Desde o início das pesquisas sérias em IA, várias décadas antes, discutia-se intensamente se os chips de silício, que codificam tudo em uns e zeros binários, poderiam realmente simular cérebros e outros processos biológicos que dão origem ao que chamamos de inteligência. Ainda não havia provas conclusivas de que isso fosse possível, e essa questão nem sequer tocava na discussão normativa sobre se *deveríamos* tentar desenvolvê-los.

Por outro lado, IA era o termo do momento para se referir tanto à tecnologia disponível quanto àquela que os pesquisadores poderiam alcançar em um futuro próximo, por meio do aperfeiçoamento dos recursos existentes. Essas ferramentas — baseadas em uma poderosa correspondência de padrões conhecida como aprendizagem de máquina — já demonstravam aplicações promissoras na mitigação das mudanças climáticas e na área da saúde.

Naquele verão, um grupo de pesquisadores, apoiado por alguns dos cientistas mais proeminentes da área, fundou uma nova organização chamada Climate Change AI, com o objetivo de incentivar o uso de técnicas e modelos de IA capazes de gerar impacto real na resolução dos desafios climáticos. Em um livro branco, a organização descreveu dez categorias de problemas especialmente adequados às capacidades atuais da aprendizagem de máquina, entre elas, tornar edifícios mais eficientes, otimizar a distribuição de carga nas redes elétricas para integrar mais energia renovável e descobrir novos materiais para geração e armazenamento de energia, além de cimento e aço mais eficientes no que se refere a carbono.

Em dezembro, a Climate Change AI organizaria um encontro lotado na NeurIPS, a conferência anual de pesquisa em inteligência artificial, um dia depois de outro grupo ter promovido, no final do mesmo corredor, um workshop igualmente concorrido, realizado em uma sala do tamanho de um campo de futebol, sobre aprendizagem de máquina aplicado à pesquisa em saúde.[5]

As palestras e os pôsteres que revestiam as paredes exibiam uma infinidade de aplicações, desde o uso da visão computacional para detectar os estágios iniciais, quase imperceptíveis, de doenças como o Alzheimer em exames de

imagem, até o uso do reconhecimento de voz para ajudar pacientes com dificuldades na fala a se comunicarem com mais facilidade.

O workshop periódico, que enfatizava a colaboração com profissionais da saúde, evoluiria dois anos depois para sua própria organização: a Machine Learning for Health.

Pesquisadores de ambas as organizações me disseram que o principal desafio de trabalhar nessas áreas não eram as limitações técnicas, era justamente o oposto. O problema estava em convencer cientistas talentosos a se dedicarem a questões que exigiam soluções relativamente simples de aprendizagem de máquina, em vez das técnicas mais avançadas que atendiam a suas ambições e valorizavam seus currículos acadêmicos. Também era preciso encontrar vontade política para implantar essas soluções em escala global.

"As tecnologias que poderiam enfrentar as mudanças climáticas estão disponíveis há anos, mas, em grande parte, não foram adotadas em escala pela sociedade", escreveram os pesquisadores da Climate Change AI em seu livro branco. Embora esperassem que a IA "fosse útil na redução dos custos associados às ações climáticas, a humanidade também deve decidir agir".[6]

De volta à sala de conferências, Sutskever interveio. Quando se tratava de resolver desafios globais complexos, "fundamentalmente, o gargalo é que você tem um grande número de seres humanos, e eles não se comunicam tão rapidamente, não trabalham tão rapidamente, têm muitos problemas de incentivo". A IAG seria diferente, afirmou ele. "Imagine uma grande rede de computadores inteligentes — todos realizando diagnósticos médicos, trocando resultados entre si com extrema rapidez."

Aquilo me soou como outra maneira de dizer que o objetivo da IAG era substituir os seres humanos. Era isso que Sutskever queria dizer? Perguntei a Brockman algumas horas depois, quando estávamos a sós.

"Não", respondeu ele rapidamente. "Isso é algo realmente importante. Qual é o propósito da tecnologia? Por que ela existe? Por que a construímos? Há milhares de anos criamos tecnologias, certo? Fazemos isso porque elas servem às pessoas. A IAG não será diferente — não da maneira como a imaginamos, não da maneira como queremos construí-la, não da maneira como acreditamos que ela deve funcionar."

Dito isso, ele reconheceu, alguns minutos depois, que a tecnologia sempre destruiu alguns empregos e criou outros. O desafio da OpenAI seria desenvolver uma IAG capaz de conceder a todos "liberdade econômica" e, ao mesmo tempo, permitir que continuassem a "viver vidas significativas" nessa nova realidade. Se fosse bem-sucedida, ela dissociaria a necessidade de trabalhar pela própria sobrevivência.

"Na verdade, acho isso algo muito bonito", disse ele.

Em nossa reunião com Sutskever, Brockman me lembrou do panorama geral.

"O que consideramos ser nosso papel não é, na verdade, determinar se a IAG será construída", afirmou.

Esse era um argumento recorrente no Vale do Silício: a inevitabilidade. Se não fizermos, alguém fará.

"A trajetória já está traçada", enfatizou, "mas o que podemos influenciar são as condições iniciais em que ela nasce."

"O que é a OpenAI?", continuou. "Qual é nosso objetivo? O que realmente estamos tentando fazer? Nossa missão é garantir que a IAG beneficie toda a humanidade, e a maneira como queremos fazer isso é construindo a IAG e distribuindo seus benefícios econômicos."

Seu tom era firme e conclusivo, como se tivesse respondido a todas as minhas perguntas. E, no entanto, de algum modo, havíamos voltado exatamente ao ponto de partida.

Minha conversa com Brockman e Sutskever se arrastou em círculos por 45 minutos, quando o tempo se esgotou. Tentei, sem muito sucesso, obter detalhes mais concretos daquilo que eles estavam tentando construir, exatamente — o que, por natureza, disseram não poder saber — e por que, se não podiam saber, estavam tão confiantes de que seria benéfico.

Em determinado momento, adotei outra estratégia, pedindo que dessem exemplos das desvantagens da tecnologia. Esse era um dos pilares da mitologia fundadora da OpenAI, a ideia de que o laboratório precisava criar uma boa IAG antes que alguém criasse uma ruim.

Brockman tentou responder: deepfakes. Segundo ele, "não está claro se o mundo fica melhor com suas aplicações".

Dei meu exemplo: já que falávamos sobre mudanças climáticas, e quanto ao impacto ambiental da própria IA? Um estudo recente[7] da Universidade de Massachusetts Amherst havia apresentado números alarmantes sobre as enormes e crescentes emissões de carbono associadas ao treinamento de modelos de IA cada vez maiores.

Isso era "inegável", segundo Sutskever, mas o retorno valeria a pena, pois a AGI iria "entre outras coisas, neutralizar especificamente o custo ambiental". Ele não chegou a dar exemplos.

"É inquestionavelmente desejável que os data centers sejam o mais ecológicos possível", acrescentou.

"Sem dúvida", retrucou Brockman.

"Os data centers são os maiores consumidores de energia, de eletricidade", continuou Sutskever, naquele momento, parecendo decidido a demonstrar que estava ciente e se importava com o tema.

Após ouvir as considerações, afirmei que consumiam 2% da energia mundial.

"O Bitcoin não consome cerca de 1%?", perguntou Brockman.

"Uau!", exclamou Sutskever, em uma súbita explosão emotiva que, àquela altura, quarenta minutos após o início da conversa, soou um tanto teatral.

Mais tarde, Sutskever se sentaria com o repórter Cade Metz, do *New York Times*, em uma conversa para o livro *Criadores de gênios*,[8] que reconta de

forma narrativa o desenvolvimento da IA. Nessa entrevista, ele diria, sem qualquer traço de ironia: "Acho bastante provável que não demore muito para que toda a superfície da Terra fique coberta por data centers e usinas de energia." Haveria, segundo ele, "um tsunami de computação… quase como um fenômeno natural". A IAG — e, portanto, os data centers necessários para sustentá-la — seria "útil demais para não existir".

Tentei, mais uma vez, obter mais detalhes.

"O que você está dizendo é que a OpenAI está fazendo uma grande aposta de que conseguirá alcançar uma IAG benéfica para combater o aquecimento global antes que o próprio processo de desenvolvê-la possa agravá-lo?"

"Eu não iria tão longe nessa linha de raciocínio", interrompeu Brockman rapidamente. "A maneira como pensamos sobre isso é a seguinte: estamos em um momento de progresso da IA. Isso é maior do que a OpenAI, certo? É da área. E acho que a sociedade está realmente se beneficiando disso."

"No dia em que anunciamos o acordo", continuou, referindo-se ao novo investimento de 1 bilhão de dólares da Microsoft, "o valor de mercado da Microsoft subiu 10 bilhões de dólares. As pessoas acreditam que há um retorno financeiro positivo mesmo em tecnologia de curto prazo."

A estratégia da OpenAI, portanto, era bastante simples, explicou ele: acompanhar esse progresso.

"Esse é o padrão que devemos realmente seguir. Devemos continuar a avançar. É assim que sabemos que estamos no caminho certo."

Mais tarde naquele dia, Brockman reiterou que o principal desafio de trabalhar na OpenAI era o fato de ninguém saber realmente como seria a IAG. Ainda assim, como pesquisadores e engenheiros, sua tarefa era continuar avançando e descobrir a forma da tecnologia passo a passo.

Ele falava igual a Michelangelo, como se a IAG já existisse dentro do bloco de mármore que esculpia. Tudo o que precisava fazer era lapidar até que ela se revelasse.

Houve então uma mudança de planos. Eu tinha um almoço marcado com os funcionários no refeitório, mas algo exigia que eu deixasse o prédio. Brockman seria meu acompanhante. Atravessamos a rua e caminhamos algumas dezenas de passos até um café ao ar livre que havia se tornado o ponto de encontro favorito dos funcionários.

Isso era algo que se tornaria recorrente durante toda a visita: andares aos quais eu não podia ter acesso, reuniões das quais não podia participar, pesquisadores lançando olhares furtivos ao chefe da comunicação a cada pausa entre frases, como para se certificarem de que não haviam violado alguma política de confidencialidade.

Um tempo depois, descobri que, após a visita, Jack Clark havia emitido um alerta incomumente severo no Slack, plataforma de comunicação interna da empresa, orientando os funcionários a não falarem comigo além das conversas autorizadas. O segurança recebeu uma foto minha, acompanhada de

instruções para ficar atento caso eu aparecesse nas instalações sem permissão.

Era um comportamento estranho por si só, e ainda mais estranho vindo de uma organização que proclamava ter o compromisso da transparência. Comecei a me perguntar o que eles estavam escondendo, se tudo o que faziam supostamente fazia parte de uma pesquisa benéfica que acabaria sendo disponibilizada ao público.

Durante o almoço, e nos dias que se seguiram, investiguei mais a fundo por que Brockman havia cofundado a OpenAI. Ele ainda era um adolescente quando se tornou obcecado pela ideia de que seria possível recriar a inteligência humana.[9] O fascínio nasceu de um artigo célebre do matemático britânico Alan Turing.

A primeira seção do texto, intitulada "O jogo da imitação", que inspiraria o título da dramatização hollywoodiana de 2014 sobre a vida de Turing, começa com uma provocação: "As máquinas podem pensar?" O artigo então define o que ficaria conhecido como teste de Turing: uma forma de medir o avanço da inteligência artificial com base na capacidade de uma máquina manter uma conversa com um ser humano sem revelar sua natureza.

Era uma história clássica entre os que trabalhavam com IA. Encantado, Brockman criou o próprio jogo baseado no teste de Turing e o publicou online, alcançando cerca de 1.500 acessos. Isso o fez se sentir extraordinário.

"Percebi que era esse tipo de coisa que eu queria fazer."

Contudo, na época, sua revelação foi prematura; a IA ainda não estava pronta para o grande público, e ele não era o tipo que se isolava em um laboratório para fazer pesquisa. Em vez disso, juntou-se à Stripe, igualmente entusiasmado com a ideia de construir uma empresa e criar produtos que pudesse colocar nas mãos de usuários reais.

Na Stripe, Brockman ganhou reputação pela lendária produtividade em programação. Era considerado um "engenheiro 10x", gíria do Vale do Silício para designar um programador capaz de resolver problemas dez vezes mais rápido do que a média. Ele, porém, não era tão habilidoso com pessoas. No cargo de diretor de tecnologia, preferia de longe programar a gerenciar. Depois de tentar ser gestor de pessoas por um tempo, procurou maneiras de se livrar das responsabilidades executivas e passar a maior parte do tempo escrevendo código.

Embora se importasse genuinamente em fazer o certo pelos colegas, cometia gafes sociais que frequentemente os deixavam constrangidos. Certa vez, convidou uma funcionária da Stripe para sair, e, logo em seguida, enviou um e-mail para toda a empresa sobre o assunto, em nome da transparência total, relembra um ex-colega.

"Acho que ele tinha boas intenções, mas foi muito estranho."

Em 2015, quando a IA começou a dar grandes saltos, Brockman se despediu da Stripe, percebendo, segundo ele, que era hora de retornar à sua ambição original. Foi melhor assim para a startup: sua relutância em gerenciar pessoas ou seguir os processos da empresa tornara-se cada vez mais

problemática à medida que a Stripe amadurecia. Em suas anotações, escreveu que faria qualquer coisa para concretizar a AGI, mesmo que isso significasse trabalhar como zelador.

Quando se casou, quatro anos depois, realizou uma cerimônia civil no escritório da OpenAI, diante de uma parede de flores personalizada com o logotipo hexagonal do laboratório. Sutskever oficiou a cerimônia. No corredor, a mão robótica usada nas pesquisas segurava as alianças, como uma sentinela de um futuro pós-apocalíptico.

"Basicamente, quero trabalhar com IAG pelo resto da minha vida", disse Brockman.

Ele abordava tudo com essa mesma intensidade. Era prático e meticuloso. Trabalhava longas horas e passava noites em claro, consumido pelas tarefas que tinha pela frente. Mesmo quando não estava trabalhando, ainda se dedicava a se aprimorar. Lia livros sobre oratória e negociação, habilidades que acreditava serem essenciais para cumprir a missão da OpenAI. Tinha certa vergonha de admitir que também lia por prazer — épicos de ficção científica, como a trilogia *O problema dos três corpos*, do escritor chinês Liu Cixin, e a série *Fundação*, de Isaac Asimov.

Para os funcionários, seu foco incansável era ao mesmo tempo uma bênção e uma maldição. Nenhum detalhe era pequeno demais para escapar à sua atenção. Se uma equipe ficava travada, ele podia permanecer sentado por horas, resolvendo pessoalmente cada obstáculo de codificação. Era o motor do progresso incessante, disposto a trabalhar dia e noite para atingir os marcos o mais rápido possível.

Ao mesmo tempo, muitos achavam que ele era detalhista em excesso, incapaz de ver o panorama geral. Podia se fixar em um problema por horas, sem considerar se o contexto havia mudado ou se aquele ainda era o problema a ser resolvido. Tinha tendência a microgerenciar, e, quando achava que o ritmo da codificação estava lento, assumia rapidamente o lugar de outros membros da equipe. As pessoas que trabalhavam com ele tinham dificuldade para acompanhar seu ritmo. Muitas acabavam esgotadas.

A OpenAI não teria chegado aonde chegou sem ele, disse um ex-engenheiro que trabalhou de perto com Brockman. Porém, se tudo dependesse apenas dele, as coisas poderiam ter dado muito errado. "Greg não tem visão. Ele não é um visionário como Sam Altman. Ele só quer um problema difícil e instigante para resolver e provar que é dez vezes mais inteligente do que qualquer outra pessoa."

Perguntei a Brockman o que o motivava.

"Quais são as chances de uma tecnologia transformadora surgir durante a vida?"

Ele estava confiante de que ele — e a equipe que havia reunido — ocupavam uma posição única para conduzir essa transformação.

"O que realmente me atrai são problemas que não teriam o mesmo desfecho se eu não participasse da resolução."

Na verdade, Brockman não queria apenas ser zelador, ele queria liderar a IAG. Transbordava a energia ansiosa de quem almejava reconhecimento histórico. Queria que, um dia, sua história fosse contada com a mesma mistura de admiração e reverência com que ele narrava as dos grandes inovadores que o precederam.

Um ano antes de nossa conversa, ele havia dito a um grupo de jovens empreendedores de tecnologia, durante um retiro exclusivo em Lake Tahoe, com uma pontada de autopiedade, que os diretores de tecnologia nunca eram lembrados.

"Citem um CTO famoso", desafiou a plateia.

Os espectadores tiveram dificuldade para responder. Brockman havia provado seu argumento.

Em 2022, ele se tornaria presidente da OpenAI.

Durante nossas conversas, Brockman insistiu que nenhuma das mudanças estruturais da OpenAI indicava uma alteração em sua missão principal. Na verdade, a sociedade limitada e o novo grupo de financiadores, segundo ele, a reforçavam.

"Conseguimos atrair investidores alinhados à missão, dispostos a priorizá-la em vez dos retornos. *Isso é algo incrível.*"

A OpenAI passou a dispor dos recursos de longo prazo necessários para seguir a chamada Lei da OpenAI. E isso, enfatizou Brockman, era imperativo. Não conseguir acompanhar o ritmo do avanço seria a verdadeira ameaça capaz de minar a missão da organização. Se o laboratório ficasse para trás, não seria o melhor, e, se não fosse o melhor, não teria esperança de mudar o curso da história em direção a sua visão de uma AGI benéfica.

Só mais tarde eu perceberia todas as implicações dessa afirmação. Foi essa suposição fundamental — a necessidade de ser o primeiro ou parecer o primeiro — que colocou em movimento todas as ações da OpenAI e suas consequências de longo alcance.

Ela impôs um relógio a cada avanço da pesquisa, regido não pela cadência de uma deliberação cuidadosa, mas pelo ritmo implacável exigido para cruzar a linha de chegada antes de qualquer outro. Essa lógica passou a justificar o consumo incomensurável de recursos pela OpenAI: tanto de computação, independentemente de seu impacto ambiental, quanto de dados, cuja acumulação não poderia ser desacelerada por questões de consentimento ou pela observância de regulamentações.

Brockman voltou a citar o aumento de 10 bilhões de dólares no valor de mercado da Microsoft.

"O que isso realmente reflete é que a IA está agregando valor real ao mundo real hoje."

Reconheceu, porém, que esse valor ainda estava concentrado em uma corporação já rica, e era por isso que a OpenAI tinha a segunda parte de sua missão: redistribuir os benefícios da AGI para todos.

"Há algum exemplo histórico de uma tecnologia cujos benefícios tenham sido distribuídos com sucesso?", perguntei.

"Bem, na verdade, acho que… é interessante olhar até para a internet como exemplo", respondeu, hesitando por um instante. "Também há problemas, certo? Sempre que você tem algo supertransformador, não é fácil descobrir como maximizar o positivo e minimizar o negativo."

"O fogo é outro exemplo", acrescentou. "Ele também tem algumas desvantagens reais. Portanto, precisamos descobrir como mantê-lo sob controle e estabelecer padrões de uso compartilhados."

"Os carros são outro bom exemplo", continuou. "Muitas pessoas têm carros, e isso beneficia muita gente, mas eles também têm desvantagens, algumas externalidades que não são necessariamente boas para o mundo", concluiu, hesitante.

"Acho que o que queremos para a IAG não é tão diferente dos aspectos positivos da internet, dos carros, do fogo. A implementação, porém, é muito diferente, porque se trata de um tipo de tecnologia completamente distinto."

Seus olhos brilharam com uma nova ideia.

"Basta olhar para os serviços essenciais. As empresas de energia e as companhias elétricas são entidades muito centralizadas que fornecem produtos de baixo custo e alta qualidade, melhorando significativamente a vida das pessoas."

Era uma boa analogia. Contudo, Brockman parecia, mais uma vez, inseguro sobre como a OpenAI realmente se transformaria em um serviço essencial. Talvez, sugeriu em voz alta, isso pudesse acontecer por meio da distribuição de uma renda básica universal, ou talvez de outra forma.

Ele então voltou à única coisa da qual parecia absolutamente certo: o compromisso da OpenAI com a redistribuição dos benefícios da IAG e a garantia da liberdade econômica para todos.

"Nós realmente acreditamos nisso."

E explicou:

"A maneira como pensamos sobre isso é a seguinte: até agora, a tecnologia tem sido algo que beneficia a todos, mas também tem um efeito real de concentração. A IAG poderia tornar isso ainda mais extremo. E se todo o valor ficasse concentrado em um único lugar? Essa é a trajetória em que estamos como sociedade, e nunca vimos esse extremo. Não acho que seja um mundo bom. Não é um mundo do qual eu queira participar. Não é um mundo que eu queira ajudar a construir."

Em fevereiro de 2020, publiquei o perfil da OpenAI na *MIT Technology Review*, baseado nas observações que fiz durante o tempo passado no escritório, em quase três dezenas de entrevistas e em alguns documentos internos.[10]

"Há um desalinhamento entre o que a empresa defende publicamente e a forma como opera a portas fechadas", escrevi. "Com o tempo, isso permitiu

que uma competitividade acirrada e uma pressão crescente por mais e mais financiamento corroessem os ideais de transparência, abertura e colaboração segundo os quais foi fundada."

Horas depois, Musk respondeu ao texto com três tuítes seguidos:

"A OpenAI deveria ser mais aberta, em minha opinião."[11]

"Não tenho controle, e tenho apenas uma visão muito limitada da OpenAI. Minha confiança em Dario em relação à segurança não é alta."

"Todas as organizações que desenvolvem IA avançada devem ser regulamentadas, incluindo a Tesla."

Logo depois, Altman enviou um e-mail aos funcionários da OpenAI:[12]

"Gostaria de compartilhar algumas reflexões sobre o artigo da *Tech Review*", escreveu. "Embora definitivamente não seja catastrófico, é claramente ruim."

Segundo Brockman, foi "uma crítica justa", reconhecendo que o artigo havia identificado uma desconexão entre a percepção pública da OpenAI e a realidade interna. Essa lacuna, porém, poderia ser amenizada não por meio de mudanças estruturais, mas com alguns ajustes na forma como a empresa se comunicava.

"É bom, não ruim, que tenhamos descoberto como ser flexíveis e nos adaptar", escreveu, referindo-se à reestruturação organizacional e ao aumento da confidencialidade, "para cumprir nossa missão à medida que aprendemos mais."

A OpenAI, recomendou ele, deveria ignorar meu artigo por enquanto e, nas semanas seguintes, começar a reforçar publicamente seu compromisso com os princípios originais, mesmo sob a nova estrutura.

"Esta também pode ser uma boa oportunidade para falar sobre a API [Interface de Programação de Aplicativos] como uma estratégia de abertura e de compartilhamento de benefícios."

E seguiu:

"A questão mais séria de todas, para mim, é que alguém vazou nossos documentos internos."

Ele explicou que uma investigação já havia sido aberta e que a empresa seria mantida informada sobre o andamento. Também sugeriu que Dario Amodei e Musk se reunissem para resolver as críticas do bilionário, que eram "suaves em comparação com outras coisas que ele já disse", mas ainda assim "uma coisa ruim de se fazer".

Para evitar qualquer mal-entendido, Altman reforçou que o trabalho de Amodei e a segurança da IA eram partes essenciais da missão da OpenAI.

"Acho que, em algum momento no futuro, devemos encontrar uma maneira de defender publicamente nossa equipe", escreveu, "mas sem dar à imprensa a briga pública que eles adorariam agora."

A OpenAI não voltaria a falar comigo pelos três anos seguintes.

CAPÍTULO 4

# SONHOS DE MODERNIDADE

Em seu livro *Poder e progresso*, Daron Acemoglu e Simon Johnson, economistas do MIT e ganhadores do Prêmio Nobel, argumentam que toda revolução tecnológica começa com uma ambição mobilizadora. É a promessa de uma tecnologia capaz de beneficiar a todos que dá início à longa jornada de reunir talentos e recursos suficientes para transformá-la em realidade. Após analisar mil anos de história da tecnologia, os autores concluem que as inovações não são inevitáveis; sua promoção depende de uma crença coletiva de que vale a pena desenvolvê-las. A ironia, observam, é que justamente por essa razão as novas tecnologias raramente geram prosperidade generalizada. Aqueles que conseguem mobilizar recursos para criá-las são os que já detêm poder e capital e que, ao transformar suas ideias em realidade, acabam impondo uma visão de mundo limitada por pontos cegos e filosofias egoístas. Somente por meio de transformações sociais profundas, ou de uma resistência organizada e poderosa, é que uma tecnologia pode deixar de enriquecer poucos para, de fato, beneficiar muitos.

Os autores citam como exemplo a invenção de uma nova descaroçadora de algodão, na década de 1790.[1] A máquina transformou o sul dos Estados Unidos no maior exportador mundial de algodão, impulsionou o crescimento econômico do país e gerou lucros extraordinários para inúmeros proprietários de terras e empresas ligadas ao setor. No entanto, o impacto real foi o de intensificar a escravidão e todo o sistema brutal de desumanização e exploração que a sustentava, um regime que perduraria por mais sete décadas. Com o aumento da produção de algodão, pessoas negras escravizadas foram forçadas a trabalhar mais horas e submetidas a formas ainda mais severas de coerção física para extrair cada grama de seu trabalho. Ao mesmo tempo, aqueles que lucravam com a descaroçadora apresentavam a invenção como algo que tornava os escravizados mais felizes. "Digo com ousadia que não há raça mais feliz e satisfeita na face da Terra", declarou um congressista da Carolina do Sul.

Essas duas características das revoluções tecnológicas — a promessa de promover o progresso e a tendência de, na prática, revertê-lo para as pessoas sem poder, especialmente as mais vulneráveis — talvez nunca tenham sido tão

verdadeiras quanto neste momento, quando nos deparamos com a inteligência artificial.

Desde sua origem, o desenvolvimento e o uso da IA têm sido impulsionados por sonhos sedutores de modernidade e moldados por uma elite restrita, dotada de dinheiro e influência, capaz de concretizar a própria visão da tecnologia. Essa visão é justamente o que levou à explosão dos custos sociais, trabalhistas e ambientais que se espalham hoje pelo mundo, sobretudo, como veremos, em muitos países do Sul Global, onde as consequências da expropriação por impérios históricos ainda se fazem sentir no atraso econômico e no enfraquecimento das instituições políticas. E, no entanto, assim como o congressista da Carolina do Sul, o Vale do Silício tem retratado as experiências daqueles que estão sendo explorados e prejudicados pela tecnologia como se essas pessoas fossem mais felizes por causa dela.

A promessa que impulsiona o desenvolvimento da IA está codificada no próprio nome da tecnologia. Em 1956, seis anos depois de Turing abrir seu artigo com a pergunta "As máquinas podem pensar?", vinte cientistas, todos homens brancos, reuniram-se na Dartmouth College para fundar uma nova disciplina dedicada ao estudo dessa questão.[2]

Vindos de campos como matemática, criptografia e ciência cognitiva, eles buscavam um termo capaz de unificá-los sob uma mesma bandeira. John McCarthy, professor da instituição e organizador do encontro, usava inicialmente a expressão *estudos sobre autômatos* para descrever a busca por máquinas capazes de apresentar comportamento automático. Quando o termo não despertou interesse suficiente, ele procurou algo mais evocativo, e acabou escolhendo a expressão *inteligência artificial*.[3]

O nome *inteligência artificial* foi, portanto, uma ferramenta de marketing desde o início, uma promessa embutida no próprio termo sobre o que a tecnologia poderia oferecer. A palavra *inteligência* soa inerentemente boa e desejável, sofisticada e impressionante; algo de que a sociedade certamente gostaria de ter mais, algo que sugeria benefícios universais.

A mudança de nome funcionou. As duas palavras despertaram imediatamente maior interesse — não apenas entre os financiadores, mas entre os cientistas, ansiosos por participar de um campo nascente com ambições tão grandiosas.

Cade Metz, cronista veterano da IA, chama essa mudança de nome de pecado original do campo: grande parte do hype e do perigo que hoje cercam a tecnologia deriva da decisão fatídica de McCarthy de associá-la a esse conceito sedutor, porém nebuloso, de "inteligência". O termo favorece antropomorfismos fáceis e exageros grandiosos sobre as capacidades da tecnologia.

Em 1958, dois anos após a fundação do campo, Frank Rosenblatt, professor da Cornell University, apresentou o Perceptron, um sistema capaz de realizar correspondências simples de padrões para distinguir cartas com base

na presença ou ausência de um pequeno quadrado impresso à esquerda ou à direita. Apesar das objeções de seu principal colaborador, Rosenblatt anunciou o sistema como algo semelhante a um cérebro humano. Chegou a afirmar que, um dia, ele seria capaz de se reproduzir e adquirir consciência. Na manhã seguinte, o *New York Times* anunciou que o Perceptron seria, no futuro, "capaz de andar, falar, ver, escrever, se reproduzir e ter consciência de sua existência".

Essa tradição de antropomorfizar persiste até hoje, reforçada por narrativas de Hollywood que fundem a ideia de "IA" com antigas representações de criações humanas que, de repente, ganham vida. Desenvolvedores de inteligência artificial falam com frequência sobre como seus sistemas "aprendem", "leem" ou "criam" como os humanos.

Esse tipo de linguagem não apenas alimentou a impressão de que as tecnologias atuais de IA são muito mais capazes do que realmente são, como se tornou uma ferramenta retórica útil para as empresas evitarem responsabilidades legais. Diversos artistas e escritores processaram desenvolvedores de IA por violação de direitos autorais, alegando que suas obras foram usadas — sem consentimento e compensação — para treinar modelos de IA. Os desenvolvedores, por sua vez, defenderam-se dizendo que isso se enquadra no uso justo, argumentando que o processo seria análogo a um ser humano se "inspirar" no trabalho de outros.

Essas analogias constantes entre IA e seres humanos também alimentaram o temor de que esse tipo de software possa, algum dia, se tornar tão capaz a ponto de nos superar, e, assim, ameaçar nossa própria existência. O medo da *super*inteligência se apoia na ideia de que a IA poderia, de algum modo, ultrapassar-nos justamente na qualidade que fez dos seres humanos a espécie dominante do planeta por dezenas de milhares de anos.

O nome *inteligência artificial* também moldou as próprias concepções do campo sobre o que essa área de estudos realmente fazia. Antes, os cientistas estavam apenas construindo máquinas para automatizar cálculos, não muito diferentes do grande aparato retratado em *O jogo da imitação*, que Turing criou para decifrar o código Enigma nazista durante a Segunda Guerra Mundial.

Dessa vez, porém, eles acreditavam estar *recriando a inteligência* — uma ideia que passaria a definir as métricas de progresso da área e, décadas depois, daria origem às ambições da OpenAI.

O problema central, porém, é que não existe uma definição cientificamente consensual de inteligência. Ao longo da história, neurocientistas, biólogos e psicólogos apresentaram diferentes explicações sobre o que ela é e por que os seres humanos parecem ter mais inteligência do que qualquer outra espécie. Talvez isso se deva ao tamanho de nosso cérebro, à capacidade de raciocinar sobre problemas complexos ou de criar modelos mentais das crenças de outras pessoas. Ao longo dos séculos, inúmeros testes foram elaborados para medir a inteligência com base nessas definições; muitos deles, porém, acabaram desacreditados e caíram em desuso devido a suas origens e aplicações questionáveis.

No início do século XIX, o craniologista estadunidense Samuel Morton mediu literalmente o tamanho de crânios humanos em uma tentativa de justificar a crença racista de que os brancos, cujos crânios ele considerava, em média, maiores, tinham inteligência superior à dos negros. Gerações posteriores de cientistas descobriram que Morton havia falsificado seus números para confirmar as próprias crenças preconcebidas e que seus dados, na verdade, não mostravam diferenças significativas entre brancos e negros. Os testes de QI também surgiram como uma forma de eliminar os "deficientes mentais" da sociedade e de justificar políticas eugênicas sob o disfarce da "objetividade" científica. Testes padronizados mais recentes, como o SAT, revelaram alta sensibilidade ao contexto socioeconômico dos candidatos, sugerindo que avaliam mais o acesso a recursos e educação do que uma capacidade inata.[4]

Em um documento publicado pela primeira vez em 2004, intitulado "O que é inteligência artificial?", McCarthy admitiu que a falta de consenso em torno da inteligência natural era inerentemente confusa para um campo que tentava recriá-la. Uma revisão de 2007 de seu artigo apresenta uma longa e sinuosa seção de perguntas e respostas, com o objetivo de abordar questões básicas para um público leigo. Ela começa assim:[5]

> P. **O que é inteligência artificial?**
>
> R. É a ciência e a engenharia de criar máquinas inteligentes, especialmente programas de computador inteligentes.
>
> P. **Sim, mas o que é inteligência?**
>
> R. Inteligência é o aspecto computacional da capacidade de atingir objetivos no mundo. Há diversos tipos e graus de inteligência em pessoas, em muitos animais e em algumas máquinas.
>
> P. **Não existe uma definição sólida de inteligência que não dependa de sua relação com a inteligência humana?**
>
> R. Ainda não. O problema é que ainda não conseguimos caracterizar, de forma geral, quais tipos de procedimentos computacionais devem ser considerados inteligentes.

Como resultado, o campo da IA passou a medir seu progresso em comparação com as capacidades humanas. As habilidades e aptidões das pessoas tornaram-se o modelo para orientar a pesquisa: a visão computacional procura recriar nossa visão; o processamento e a geração de linguagem natural, nossa capacidade de ler e escrever; o reconhecimento e a síntese de fala, nossa habilidade de ouvir e falar; e a geração de imagens e vídeos, nossa criatividade e imaginação. À medida que os softwares em cada uma dessas áreas evoluiu, os pesquisadores começaram a integrá-los nos chamados sistemas multimodais

— sistemas capazes de "ver" e "falar", "ouvir" e "ler". O fato de a tecnologia ameaçar substituir uma grande parcela dos trabalhadores humanos não é coincidência, mas resultado de um projeto deliberado.

Ainda assim, a busca pela inteligência artificial segue sem um rumo definido. A cada novo avanço na pesquisa em IA, surgem debates intensos sobre se esse avanço representa uma recriação genuína da inteligência ou apenas uma imitação inferior dela. Para tentar diferenciar as duas, o termo *inteligência artificial geral* passou a ser usado como referência à inteligência verdadeira. No entanto, essa nova formulação não alterou o fato de que ainda não existe uma maneira clara de medir o progresso nem de determinar quando o campo poderá ser considerado bem-sucedido. Entre os pesquisadores, é comum dizer que aquilo que hoje é visto como IA, amanhã deixará de ser.

O teste de Turing não permaneceu por muito tempo como indicador de inteligência artificial, pois foi rapidamente superado, e os cientistas perceberam que ele não havia cumprido, de fato, seu propósito. Houve também uma época em que se acreditava que um computador capaz de vencer humanos no xadrez ou no Go seria a prova definitiva de sucesso. Hoje, o AlphaGo, da DeepMind, é visto como uma demonstração impressionante do potencial do software, porém, mais uma vez, está longe de representar a conclusão das ambições da área. Ao longo de décadas de pesquisa, a definição de IA mudou conforme seus parâmetros de referência foram sendo ajustados, reescritos e, muitas vezes, abandonados. Os objetivos do desenvolvimento da inteligência artificial permanecem em constante mutação e, como observou a diretora de pesquisa da Data & Society, Jenna Burrell, são um "horizonte do futuro cada vez mais distante".[6] O avanço da tecnologia segue em direção a um objetivo indefinido, sem um fim previsível à vista.

Para justificar o prolongamento dos prazos e os custos cada vez mais elevados na busca ambiciosa da IA, as promessas associadas a ela tornaram-se mais grandiosas do que nunca. A inteligência artificial, que antes era apenas um fascínio científico e uma tecnologia com algum potencial de aplicação comercial, passou a ser descrita como a precursora da quarta revolução industrial, a pedra angular da superpotência moderna. A IAG, se algum dia for alcançada, promete resolver as mudanças climáticas, tornar os cuidados de saúde acessíveis e proporcionar uma educação equitativa. A OpenAI é o exemplo mais emblemático desse tipo de pensamento. Ela não consegue explicar como a tecnologia cumprirá tais promessas — apenas afirma que o preço exorbitante que a sociedade está pagando por seu desenvolvimento valerá a pena algum dia.

O que não se diz é que, na ausência de um significado consensual, "inteligência artificial", ou mesmo "inteligência artificial geral", pode significar simplesmente aquilo que a OpenAI desejar que signifique.

A história da IA revela que seu desenvolvimento sempre foi moldado por uma elite poderosa. Não é por acaso que, hoje, a inteligência artificial se tornou

sinônimo de modelos colossais e altamente dependentes de recursos, tecnologias que apenas um pequeno grupo de empresas tem capacidade de desenvolver e que desejam transformar em alicerce para tudo. Mesmo em seus primórdios, antes que os interesses comerciais tornassem a política da revolução da IA mais evidente, as explorações científicas do campo já oscilavam e se desviavam em meio a intensas disputas por financiamento e influência.

Após o encontro em Dartmouth, formaram-se dois campos com teorias concorrentes sobre como avançar nessa área. O primeiro, conhecido como simbolista, acreditava que a inteligência deriva do conhecimento: os seres humanos sabem mais do que os animais e são capazes de usar esse conhecimento para compreender e agir no mundo. Portanto, alcançar a IA deveria envolver a codificação, nas máquinas, de representações simbólicas do conhecimento sobre o mundo, dando origem aos chamados sistemas especialistas.

O segundo grupo, conhecido como conexionista, defendia que a inteligência surge da aprendizagem. Os seres humanos têm uma capacidade de aprender superior à dos animais e podem usar essa habilidade para adquirir e aperfeiçoar diferentes competências. Assim, o desenvolvimento da IA deveria concentrar-se na criação dos chamados sistemas de aprendizagem de máquina, que buscariam imitar a forma como o cérebro humano processa sinais e informações. Essa hipótese acabaria conduzindo à popularização das redes neurais, softwares de processamento de dados projetados, ainda que de forma aproximada, para refletir as conexões interligadas do cérebro, que hoje constituem a base da IA moderna, incluindo todos os sistemas de IA generativa.

Nas décadas seguintes, os dois campos disputaram um financiamento limitado e o controle sobre a imaginação popular acerca do que a IA poderia ser. Naquele período, essas rivalidades se desenrolavam principalmente em universidades e revistas acadêmicas, entre cientistas que competiam por recursos do governo e de fundações; ocasionalmente, seus debates transbordavam para a mídia, moldando a percepção pública sobre o tema.

À frente dos conexionistas estava Frank Rosenblatt e seu Perceptron, uma das primeiras provas de conceito de um sistema de aprendizagem de máquina. Rosenblatt não forneceu instruções explícitas ao sistema, projetou-o para que calculasse as próprias regras ao aprender a distinguir diferentes cartões depois de exposto a inúmeros exemplos.

À frente dos simbolistas estava o arqui-inimigo de Rosenblatt, o professor Marvin Minsky, do MIT. Coorganizador do workshop de Dartmouth, Minsky havia se interessado pelo pensamento conexionista, mas acabou se desiludindo com a ideia. E não mudou de posição de forma discreta. Aproveitou inúmeras oportunidades para se exibir e ridicularizar os colegas conexionistas que competiam com ele pelas mesmas bolsas, chegando até a atacar pesquisadores em início de carreira. Em 1969, assinou, com outros

autores, o livro *Perceptrons*, cuja crítica contundente ao conexionismo é amplamente considerada, junto ao avanço limitado das redes neurais na época, responsável por praticamente eliminar o financiamento dessa linha de pesquisa por mais de quinze anos.[7]

Em sua disputa por influência, Minsky contava com um aliado improvável: Joseph Weizenbaum, seu colega do MIT.[8] Em 1958, mesmo ano em que o Perceptron de Rosenblatt encantou o *New York Times*, Weizenbaum criou um sistema de IA chamado ELIZA, que também ganhou destaque nas manchetes. Mais de seis décadas antes do ChatGPT, o ELIZA era um chatbot com o qual se podia conversar por meio de uma máquina de escrever elétrica conectada a um enorme mainframe que ocupava uma sala inteira. No entanto, enquanto o ChatGPT surgiu da tradição conexionista, o ELIZA era um sistema simbolista, que utilizava regras simples para parecer atencioso e responsivo ao usuário.

Uma das versões mais populares do ELIZA foi inspirada em um método de terapia conversacional que explorava as falas do paciente, refletindo o que ele dizia e pedindo, de modo gentil, mais informações. Em um artigo[9] que Weizenbaum publicou posteriormente sobre o funcionamento interno do programa, ele apresentou o seguinte diálogo como exemplo:

**USUÁRIO**: Os homens são todos iguais.

**ELIZA**: DE QUE MANEIRA

**USUÁRIO**: Eles estão sempre nos incomodando com algo.

**ELIZA:** VOCÊ PODE ME DAR UM EXEMPLO ESPECÍFICO

**USUÁRIO**: Bem, meu namorado me fez vir aqui.

**ELIZA**: SEU NAMORADO FEZ VOCÊ VIR AQUI

**USUÁRIO**: Ele diz que estou deprimida a maior parte do tempo.

**ELIZA**: LAMENTO SABER QUE VOCÊ ESTÁ DEPRIMIDA

**USUÁRIO**: É verdade. Estou infeliz.

**ELIZA**: VOCÊ ACHA QUE VIR AQUI PODE AJUDÁ-LA A DEIXAR DE SER INFELIZ

Weizenbaum projetou o sistema como um experimento para testar quão facilmente os seres humanos poderiam ser enganados pela ilusão de inteligência. O nome ELIZA foi inspirado em Eliza Doolittle, personagem da classe trabalhadora interpretada por Audrey Hepburn no filme *Minha bela dama*, de 1964, que aprende a se passar por uma duquesa da alta sociedade após ser ensinada por um homem rico a mudar a dicção e os modos.

O sucesso subsequente do ELIZA[10] em convencer pessoas de que era

inteligente alarmou Weizenbaum. A demonstração pareceu tão convincente para alguns que psiquiatras começaram a discutir a psicoterapia automatizada como algo iminente. Apenas alguns anos após a fundação do campo da IA, cientistas da computação já concluíam, de forma prematura, que a compreensão da linguagem natural pelos computadores era um problema resolvido. (Décadas depois, se isso realmente se confirmou ou não continua sendo uma questão em aberto.)

Weizenbaum passou boa parte do restante da carreira tentando conter o entusiasmo em torno de sua criação e combatendo a própria premissa fundamental que sustentava a busca pela IA. Ele enfatizava que o ELIZA não passava de um programa puramente procedural, codificado para identificar palavras-chave na fala do usuário e aplicar transformações simples para gerar respostas. Assim, "meu namorado" tornava-se "seu namorado"; "estou deprimida" transformava-se em "você está deprimida".

Na realidade, não havia algo de verdadeiramente inteligente naquele sistema. Anos depois, Weizenbaum publicou o livro *Computer Power and Human Reason* [O poder dos computadores e a razão humana, em tradução livre], lançado na década seguinte a *Perceptrons*, de Minsky.[11] Nele, argumentava que seres humanos e máquinas são fundamentalmente diferentes e que a tentativa da IA de dissolver essa distinção traria consequências sociais profundas. Isso abriria caminho, por exemplo, para que pessoas no poder — sejam CEOs ou políticos — impusessem sua vontade por meio das máquinas, ao mesmo tempo que se eximiriam de qualquer responsabilidade moral.

Apesar dos esforços de Weizenbaum, a impressionante demonstração do sistema simbólico ELIZA acabou, inadvertidamente, reforçando a campanha de Minsky para consolidar o simbolismo como a vertente dominante sobre o conexionismo. Nas décadas seguintes, até os anos 1990, os sistemas especialistas tornaram-se o foco principal da pesquisa e da comercialização de IA.

Esse pensamento dominante levou a projetos como o Cyc, uma iniciativa que buscava desenvolver um sistema de senso comum ao programá-lo com 100 milhões de regras sobre a vida cotidiana. Contudo, em diversos momentos, os avanços nos sistemas simbólicos começaram a estagnar, à medida que os esforços para ampliá-los esbarravam no enorme desafio de codificar manualmente todas as regras. Afinal, como registrar todas as sutilezas da língua inglesa, gírias, sarcasmos, figuras de linguagem e exceções gramaticais?

Cada vez que os obstáculos se multiplicavam, os financiadores perdiam o interesse, mergulhando o campo em um estado de crise existencial que ficaria conhecido como o "inverno da IA".[12]

Neste ponto, a história da IA costuma ser narrada como um triunfo do mérito científico sobre a política. Minsky pode ter usado sua reputação e influência para suprimir o conexionismo, mas a força da ideia acabou prevalecendo,

permitindo que essa vertente ocupasse seu lugar de direito como a base da revolução moderna da IA.

Durante o período em que o simbolismo dominava, um pequeno grupo de conexionistas permaneceu fiel à busca iniciada por Rosenblatt por sistemas de aprendizagem de máquina e continuou a desenvolvê-los. Entre eles estava Geoffrey Hinton, orientador de doutorado de Ilya Sutskever, que, como professor na Universidade Carnegie Mellon na década de 1980, realizou, com colegas da UCSD, em San Diego, uma melhoria fundamental nas primeiras redes neurais.

Naquela época, os conexionistas levantaram a hipótese de que suas redes neurais fracassavam por serem muito simples, compostas por apenas uma única camada de "neurônios" interconectados, ou nós de processamento de dados. Para imitar de forma mais fiel o funcionamento do cérebro humano, o software provavelmente precisava de várias camadas empilhadas e conectadas, formando uma rede neural profunda.

Geoffrey Hinton e seus coautores tornaram essa transformação possível ao empregar um algoritmo conhecido como retropropagação, que permite que redes neurais profundas troquem e processem informações entre suas camadas. Mais tarde, em um movimento de rebranding, Hinton batizou esse processamento multicamadas de deep learning, abreviação para o uso de redes neurais *profundas* na realização de aprendizagem de máquina.

Contudo, as redes neurais profundas — hoje conhecidas como redes neurais — surgiram décadas antes de seu tempo. Para realmente demonstrarem seu potencial, precisavam de mais poder de processamento do que os computadores da década de 1980 podiam oferecer e de muito mais exemplos, ou dados, do que era possível reunir de forma acessível no mundo analógico. Em sua essência, as redes neurais são calculadoras estatísticas que identificam padrões em dados existentes — textos, imagens ou vídeos — e os utilizam para interpretar ou gerar novos dados.

Hoje, se um desenvolvedor de IA quiser criar um modelo capaz de detectar pessoas em imagens, ele pode alimentar uma rede neural com centenas de milhares de imagens, cada uma rotulada com um valor — 1 para "tem uma pessoa" e 0 para "não tem uma pessoa". (Quando você resolve captchas do Google clicando em todas as imagens com faixa de pedestre, na verdade está ajudando a treinar as redes neurais da empresa.) Usando métodos estatísticos, a rede neural identifica padrões de pixels nas imagens que correspondem à presença ou à ausência de uma pessoa. Esse processo é conhecido como treinamento de um modelo de IA. Depois que o modelo é treinado, o desenvolvedor pode executá-lo em novos dados — processo chamado de inferência — para determinar se eles se encaixam no padrão aprendido. Em outras palavras: esta imagem é um 1 ou um 0? Ela contém uma pessoa ou não?

De modo geral, as redes neurais precisam ser treinadas com um limite de dados de alta qualidade e um nível adequado de poder computacional para

identificar padrões e gerar um modelo de IA com bom desempenho. Hinton e seus colegas estavam à frente de seu tempo.

No entanto, no fim dos anos 2000, com o avanço dos computadores e o amadurecimento da internet, que passou a oferecer vastos repositórios de dados digitais, as redes neurais finalmente encontraram as condições ideais para prosperar. Pouco tempo depois, o Google adquiriu a DNNresearch, empresa fundada por Hinton (a sigla DNN significa Deep Neural Networks, ou redes neurais profundas), marcando o início de uma nova corrida pela comercialização do deep learning.

Nessa narrativa, a lição aprendida é: a ciência é um processo confuso, mas, no fim das contas, as melhores ideias acabam prevalecendo, mesmo diante dos críticos mais ferrenhos. Há também uma mensagem implícita: a tecnologia avança em um movimento inevitável de progresso.

No entanto, há outra forma de interpretar essa história. O conexionismo não superou o simbolismo apenas por mérito científico; ele também atraiu o apoio de financiadores poderosos graças a vantagens estratégicas que despertaram o interesse comercial desses investidores.

A força da IA simbólica reside na codificação explícita de informações e das relações entre elas dentro do sistema, o que permite recuperar respostas precisas e realizar raciocínios, uma característica da inteligência humana considerada essencial para sua replicação. Um exemplo emblemático é o Watson, da IBM, um dos sistemas simbólicos mais famosos, que impressionou o público no programa *Jeopardy!* em 2011. A rapidez em fornecer respostas corretas durante o jogo baseava-se na capacidade de vasculhar vastos bancos de conhecimento e reproduzi-los com precisão, demonstrando o poder e as limitações dessa abordagem.

Por outro lado, a fraqueza do simbolismo mostrou-se um obstáculo significativo: sua comercialização era lenta, cara e imprevisível. Após lançar o Watson em um programa de TV noturno, a IBM logo descobriu que transformar o sistema em algo capaz de gerar os tipos de resultados pelos quais os clientes realmente pagariam, como responder a perguntas médicas complexas, em vez de triviais, exigiria anos de investimento inicial, sem qualquer garantia de retorno. Depois de gastar mais de 4 bilhões de dólares sem perspectivas concretas de lucro, a IBM acabou abandonando o projeto e vendendo o Watson Health em 2022 por um quarto desse valor.

As redes neurais, por sua vez, apresentam vantagens e desvantagens de outra natureza. Durante anos, o campo debateu intensamente se esse tipo de software conexionista seria capaz de fazer o que os sistemas simbólicos fazem: armazenar informações e raciocinar. Independentemente da resposta, ficou evidente que, se as redes neurais conseguem fazê-lo, o fazem de forma ineficiente. Apenas com quantidades extraordinárias de dados e enorme poder computacional elas começaram a demonstrar comportamentos que sugerem o surgimento de qualquer uma dessas capacidades.

Dito isso, há um aspecto em que os modelos de deep learning realmente se

destacam: a facilidade de comercialização. Não é necessário que sejam sistemas perfeitamente precisos ou dotados de raciocínio complexo para gerar lucros expressivos. A simples capacidade de identificar e prever padrões estatísticos com alta correlação já é suficiente para resolver problemas altamente lucrativos. Além disso, o retorno financeiro tende a ser rápido e previsível, mesmo diante de investimentos iniciais elevados, algo perfeitamente alinhado aos ciclos de planejamento corporativo e ao ritmo dos lucros trimestrais. Melhor ainda, esses modelos podem ser reaplicados em diversos contextos sem exigir conhecimento técnico especializado, o que os torna ideais para empresas de tecnologia com ambições de crescimento. E, por fim, a deep learning oferece uma vantagem competitiva decisiva: favorece os participantes que detêm os maiores volumes de dados.

As gigantes da tecnologia já estavam vendo os primeiros sinais do potencial comercial das redes neurais antes do leilão da DNNresearch.[13] Em 2009, os alunos de pós-graduação de Hinton mostraram que esse software era bom para reconhecimento de voz. A IBM, a Microsoft e o Google aderiram à tendência, mas o Google foi o mais rápido a alcançar a comercialização. Em 2012, colocou as redes neurais em produção, melhorando significativamente os recursos de reconhecimento de voz do Android, ao mesmo tempo que mais alunos de pós-graduação de Hinton, desta vez Sutskever e Alex Krizhevsky, alcançaram resultados inovadores na ImageNet, demonstrando que as redes neurais também eram muito boas no reconhecimento de imagens.

O bem-sucedido lançamento do Android preparou o terreno para que o Google estivesse disposto a investir pesado nos três acadêmicos, marcando o início da adoção completa da deep learning pela indústria de tecnologia.

Hinton, Sutskever e Krizhevsky continuaram a promover o uso das redes neurais dentro do Google, ganhando força ao aplicar essa tecnologia a uma ampla gama de problemas técnicos de alto valor comercial. Trabalhando em paralelo, eles desenvolveram modelos de deep learning para tradução automática, aprimorando o Google Tradutor; para previsão de texto, introduzindo o recurso de sugestões automáticas de conclusão no Gmail; e para um projeto ambicioso de carro autônomo, que mais tarde se tornaria a Waymo.

À medida que as operações de IA do Google se expandiam, as redes neurais começaram também a gerar melhorias decisivas em sua principal fonte de receita: o mecanismo de busca. O software passou a correlacionar com mais precisão as consultas dos usuários com páginas da web relevantes, oferecendo resultados de maior qualidade e, mais importante, anúncios mais direcionados.

Quanto mais o Google lucrava e investia bilhões no deep learning, mais o restante da indústria seguia o mesmo caminho. As empresas de tecnologia rapidamente ultrapassaram governos e fundações como os principais financiadores da pesquisa em IA, e logo passaram a ditar a agenda científica com base em avanços que também prometiam retorno financeiro de curto

prazo.

A combinação entre a deep learning e os interesses comerciais transformou simultaneamente a indústria de tecnologia e o próprio desenvolvimento da inteligência artificial. Para o público, a IA generativa pareceu surgir do nada no fim de 2022, com o lançamento do ChatGPT pela OpenAI. No entanto, entre 2012 e 2022, a começar com o avanço da ImageNet, foram as transformações ocorridas durante a primeira grande era de comercialização da IA que lançaram as bases para muitas das características da revolução generativa que vemos hoje.

Para a indústria, a deep learning impulsionou a criação e o aperfeiçoamento de produtos e serviços, desde o acesso mais rápido à informação e o comércio eletrônico mais eficiente até o surgimento da economia compartilhada. Para a deep learning, a indústria impulsionou avanços técnicos em redes neurais e chips de computador que permitiram o desenvolvimento de modelos de IA maiores e mais poderosos.

Contudo, paralelamente a esses avanços impressionantes, o impulso dado pela deep learning ao Vale do Silício também ampliou seu modelo de negócios, aquele que a professora de Harvard Shoshana Zuboff batizou, em 2014, como *capitalismo de vigilância*.[14] Enquanto o capitalismo industrial extraía valor da produção de bens materiais que as pessoas desejavam comprar, o capitalismo de vigilância, argumentava Zuboff, transformava os próprios usuários em produtos. As grandes empresas de tecnologia, detentoras de enormes quantidades de dados pessoais, passaram a alimentar redes neurais com esses dados, permitindo traçar perfis de usuários com precisão sem precedentes e explorar seu engajamento para gerar receita publicitária. Para superar a concorrência, essas empresas buscaram coletar ainda mais dados, registrando logs cada vez mais detalhados de cliques, rolagens e curtidas e incentivando os usuários a fornecerem artefatos digitais cada vez mais íntimos, todas as trocas de e-mail, as fotos dos filhos, as conversas com amigos e até os pensamentos sobre questões sociais e políticas.

Ao mesmo tempo, o impulso que o Vale do Silício deu à deep learning em sua busca por expandir e consolidar monopólios em escala global também criou, entre os desenvolvedores de IA, uma cultura que passou a enxergar tudo como dado a ser capturado e consumido por suas tecnologias, em uma nobre tentativa de fazer com que representassem o mundo o mais fielmente possível. Em 2023, um grupo de pesquisadores em IA, incluindo Ria Kalluri, da Universidade Stanford, William Agnew, da Universidade de Washington, e Abeba Birhane, da Fundação Mozilla, analisaria mais de quarenta mil artigos e patentes sobre visão computacional e observaria o uso predominante de uma linguagem abstrata e distante a fim de tornar mais aceitável e normalizar a dependência que a área tem da raspagem e extração de dados em massa.[15]

Os rastros digitais detalhados dos pensamentos e ideias das pessoas nas

redes sociais eram meramente "texto". Pessoas e veículos nas fotos eram meramente "objetos". A vigilância era meramente "detecção".

Essa cultura está no centro de um intenso debate sobre a IA generativa, em especial sobre se as empresas de tecnologia têm o direito de coletar livros e obras de arte em massa para treinar seus sistemas. Para muitos desenvolvedores de inteligência artificial, que há anos operam sob essa mentalidade, a questão soa quase estranha; tratá-la com seriedade é visto como um obstáculo direto à busca moral pelo progresso contínuo. Ainda que alguns desses desenvolvedores tenham se tornado mais conscientes e inquietos diante do abismo entre sua perspectiva e a dos autores e artistas que se opõem a essa prática, essa forma de pensar permanece profundamente enraizada e tem se mostrado difícil de transformar.

Em maio de 2023, pouco depois de um grupo de artistas entrar com o primeiro processo judicial contra desenvolvedores de IAG por uso indevido de suas obras de arte, estive em Ruanda, cobrindo uma conferência de pesquisa em inteligência artificial como repórter do *Wall Street Journal*. Enquanto caminhava pelo salão abobadado do reluzente centro de convenções em forma de cúpula na capital do país, uma pesquisadora sênior me abordou e perguntou se o *WSJ* impresso no meu crachá se referia a uma nova startup ou à publicação jornalística. Quando esclareci que era jornalista e que realmente se tratava do *Wall Street Journal*, outra pesquisadora sênior se intrometeu:

"Eu reconheci por causa do conjunto de dados do WSJ", disse ela, referindo-se a um dos primeiros datasets de reconhecimento de voz por IA, composto por meio da leitura que pessoas faziam de trechos do jornal. "Já trabalhei com ele muitas vezes."

Encontrei-me presa a esse mesmo modo de pensar quando comecei a cobrir o tema da IA em 2018. Após assimilar a linguagem da comunidade para dialogar com pesquisadores da área, fiquei fascinada pelas inúmeras formas como eles extraíam e produziam conjuntos de dados. Em um exemplo que me pareceu particularmente engenhoso, pesquisadores utilizaram milhares de vídeos postados no YouTube do Desafio do Manequim, de 2016, no qual pessoas permaneciam imóveis enquanto as câmeras faziam panorâmicas e zooms ao redor delas, para treinar modelos de IA capazes de processar cenas tridimensionais.

Em 2019, uma investigação da NBC[16] conduzida por Olivia Solon acabou com minha visão romântica. Solon revelou que um software de reconhecimento facial havia sido treinado com milhões de fotos pessoais do Flickr, coletadas sem o consentimento dos usuários. O que mais me surpreendeu não foram as descobertas em si — eu já sabia havia muito tempo que o Flickr era uma das fontes de dados preferidas dos pesquisadores de IA. O que realmente me chocou foi perceber o quanto eu mesma havia passado a encarar essa prática como algo completamente normal.

Com essa nova consciência, comecei a perceber como a pressão incessante por mais dados de treinamento estava impulsionando uma vigilância

generalizada não apenas no mundo digital, mas também no mundo físico. Notei, ainda, como o olhar dessa vigilância recaía repetidamente sobre populações vulneráveis, como crianças e grupos historicamente marginalizados, especialmente nos países em desenvolvimento.[17]

Naquele mesmo ano, deparei-me com uma startup sediada em Massachusetts e incubada pela Harvard que comercializava faixas de cabeça com tecnologia de IA, supostamente capazes de medir a atividade das ondas cerebrais dos alunos para informar aos professores se os jovens estavam concentrados ou não. A empresa testava o produto em escolas primárias da Colômbia e da China em troca do direito de utilizar os dados dos estudantes para aperfeiçoar a própria tecnologia.

"Temos a vantagem de sermos os primeiros", afirmou um pesquisador científico da empresa durante uma conferência sobre tecnologia educacional em 2017. "Seremos capazes de construir um dos maiores bancos de dados de ondas cerebrais do mundo. Todos esses dados nos ajudarão a aprimorar nossos algoritmos e, consequentemente, nossos produtos, criando uma barreira à entrada ainda maior."[18]

Alguns meses depois de eu conhecer a startup, um clamor público por privacidade de dados na China, liderado por pais indignados com a ideia de ver os filhos transformados em cobaias, forçou a empresa a mudar o foco de aplicação de sua tecnologia.[19] Ainda assim, o episódio me deixou uma sensação incômoda, a de que aquela revolta bem-sucedida era a exceção, não a regra. A abordagem original da empresa — buscar países ansiosos por abraçar a promessa da tecnologia para encontrar doadores de dados e testadores de produtos — parecia, na verdade, estar se tornando uma tendência global.

Quando compartilhei essa preocupação com uma colega, ela me apresentou uma expressão já cunhada para descrever esse fenômeno: "colonialismo de dados". Foi então que descobri o trabalho dos pesquisadores Nick Couldry e Ulises A. Mejias,[20] cujo livro fundamental, *The Costs of Connection* [Os custos da conexão, em tradução livre], publicado naquele mesmo ano, argumentava que a dataficação generalizada de tudo promovida pelo Vale do Silício estava provocando o retorno de padrões históricos de conquista e extrativismo.

No ano seguinte, o artigo "Decolonial AI" [IA Decolonial, em tradução livre], escrito por Shakir Mohamed e William Isaac[21] (ambos do DeepMind) em colaboração com Marie-Therese Png, da Universidade de Oxford, reforçou uma suspeita que eu já vinha alimentando: a indústria de IA, ao mesmo tempo dependente e propulsora da dataficação, estava acelerando a reprodução desse novo tipo de colonialismo.*

Pouco tempo depois, em 2021,[22] observei a mesma dinâmica da startup de educação em IA se repetir na África do Sul. Empresas de reconhecimento facial de diferentes partes do mundo competiam para coletar dados valiosos de rostos no país, especialmente após o setor enfrentar fortes críticas por falhas em identificar com precisão pessoas de pele mais escura. Durante essa

investigação, conheci Thami Nkosi, um ativista local nascido e criado em um dos bairros mais pobres de Joanesburgo, uma área que outrora servira como depósito de resíduos químicos da indústria de mineração. Ele me mostrou as milhares de câmeras espalhadas pelas ruas da cidade e explicou como elas restringiam a liberdade de movimento da população negra, já oprimida pelo legado racial do Apartheid e temerosa de ser criminalizada simplesmente por existir em um bairro branco.

"Basicamente, eles estão monetizando os espaços públicos e a vida pública", afirmou Nkosi.

Com uma clareza cada vez maior, percebi que a própria revolução tecnológica que prometia um futuro melhor para todos estava, na verdade, revivendo, para as pessoas à margem da sociedade, os aspectos mais sombrios do passado.

Contudo, mesmo que a visão de IA do Vale do Silício revelasse os próprios desafios, a primeira era de comercialização da IA também sufocou alternativas. À medida que as grandes empresas injetavam somas sem precedentes em deep learning e no conexionismo, eclipsando praticamente todas as outras fontes de financiamento, elas reconfiguraram o panorama da pesquisa de acordo com as próprias prioridades comerciais.

Entre 2013 e 2022,[23] os investimentos da iniciativa privada em IA, incluindo fusões e aquisições, saltaram de 14,6 bilhões de dólares para 235 bilhões de dólares, atingindo um pico de 337,4 bilhões de dólares em 2021, segundo o Índice de IA da Universidade Stanford. E esses números nem sequer incluem os gastos internos das empresas com pesquisa e desenvolvimento. Em 2021, a Alphabet e a Meta gastaram, respectivamente, 31,6 bilhões e 24,7 bilhões de dólares.[24] Em contraste, o governo dos Estados Unidos destinou apenas 1,5 bilhão de dólares naquele ano ao desenvolvimento de IA não relacionada à defesa, enquanto a Comissão Europeia alocou cerca de 1 bilhão de euros.[25]

Os talentos naturalmente seguiram o dinheiro. Muitos professores universitários reformularam linhas de pesquisa em torno das redes neurais, atraídos tanto pelos resultados expressivos quanto pelo maior acesso a financiamento das corporações. O mesmo ocorreu com inúmeros estudantes de graduação e pós-graduação, motivados pela segurança de carreira oferecida pela deep learning e pela escassez de oportunidades viáveis em outros métodos de IA. As empresas de tecnologia, por sua vez, passaram a estreitar laços com a academia, firmando acordos de cooperação que aprofundaram a integração entre pesquisa acadêmica e interesses empresariais.

Em 2013, Geoffrey Hinton ingressou no Google, sob a condição de manter o cargo na Universidade de Toronto. No ano seguinte, o Facebook firmou um acordo semelhante com Yann LeCun, que concluiu o pós-doutorado em Hinton e era professor da Universidade de Nova York. Mais tarde, ambos dividiram o Prêmio Turing de 2018, frequentemente chamado

de “Prêmio Nobel da Computação”, com Yoshua Bengio, professor da Universidade de Montreal, em reconhecimento ao trabalho pioneiro dos três no deep learning. A premiação consagrou o trio como os “padrinhos da IA”. Hinton ainda viria a receber um Prêmio Nobel real em 2024, junto a outro cientista. Seguindo o exemplo de Hinton e LeCun, muitos professores e pesquisadores de IA passaram a manter afiliações duplas, com empresas de tecnologia e universidades ao mesmo tempo. Em larga escala, essa prática começou a corroer a independência científica.

Cada vez mais, pesquisadores também começaram a abandonar completamente a academia. Entre 2006 e 2020, o êxodo para a indústria entre professores pesquisadores em IA aumentou oito vezes; e, entre 2004 e 2020, a proporção de doutores em IA que seguiram carreira em empresas saltou de 21% para 70%, segundo um estudo publicado em 2023 na revista *Science* por pesquisadores do MIT.[26] Muitos foram atraídos inicialmente pela remuneração astronômica, que, para especialistas experientes, podia chegar a 1 milhão de dólares por ano.[27] Em 2015, por exemplo, a Uber contratou quarenta dos cem pesquisadores de IA de um único laboratório da Universidade Carnegie Mellon, depois de se instalar na cidade e oferecer salários que chegavam até o dobro daqueles pagos pela universidade.[28]

Com o tempo, porém, outro fator passou a pesar ainda mais: o custo crescente da pesquisa em deep learning. As universidades deixaram de ter condições de arcar com os gastos de chips de processamento e energia elétrica necessários para desenvolver trabalhos nas áreas mais promissoras da IA. Como consequência, o mesmo estudo da *Science* revelou que, em apenas três anos, de 2017 a 2020, a proporção de modelos de IA de melhor desempenho vinculados à indústria saltou de 62% para impressionantes 91%.

Na metade da primeira década de comercialização da IA, a maior parte da pesquisa de ponta já estava sendo conduzida em laboratórios acadêmicos vinculados a grandes empresas de tecnologia. Em um estudo,[29] Ria Kalluri, William Agnew, Abeba Birhane e outros colaboradores constataram que 55% dos artigos de pesquisa mais influentes sobre IA publicados em 2018 e 2019 tinham pelo menos um coautor do setor privado, um salto em relação aos 24% registrados uma década antes.

A pesquisa também passou a se concentrar em poucas empresas. No mesmo período, gigantes como Microsoft e Google mais que triplicaram a participação em publicações vinculadas à indústria, alcançando 66% do total. Ironicamente, foi exatamente essa a razão que Elon Musk e Sam Altman apontaram para fundar a OpenAI. O lucro da indústria de tecnologia havia se tornado a força dominante que impulsionava o desenvolvimento da inteligência artificial.

O impacto dessa concentração de financiamento e talentos durante a primeira era da IA reduziu drasticamente a diversidade de ideias dentro da pesquisa. O deep learning manteve-se como o paradigma dominante não apenas pelo

mérito científico, mas também pela existência de pouquíssimo investimento na exploração e no desenvolvimento de abordagens alternativas. De fato, embora as redes neurais sejam criações notáveis, com inúmeras aplicações relevantes, suas limitações — em especial, suas formas contestadas e ineficientes de armazenar informações precisas e realizar raciocínio — permaneceram evidentes, à medida que as empresas seguiram expandindo seu uso para uma lista cada vez maior de contextos e aplicações.

As redes neurais demonstraram, por exemplo, ser pouco confiáveis e imprevisíveis. Como comparadores de padrões estatísticos, elas às vezes se concentram em padrões estranhamente específicos ou completamente equivocados. Um modelo de deep learning pode reconhecer pedestres apenas pelas faixas de pedestres sob eles, deixando de identificar uma pessoa que atravessa a rua fora da faixa. Pode também associar uma placa de pare ao fato de estar à beira da estrada e deixar de reconhecer a mesma placa na lateral de um ônibus escolar ou sendo segurada por um guarda de trânsito.[30]

As redes neurais são ainda altamente sensíveis a alterações em seus dados de treinamento. Alimente-as com um conjunto diferente de imagens de pedestres ou de semáforos, e elas aprenderão um conjunto completamente novo de associações. O problema é que essas mudanças são incompreensíveis. Ao abrir o capô de um modelo de deep learning, o que se encontra são cadeias de números altamente abstratas. É isso que os pesquisadores querem dizer ao chamar o deep learning de "caixa-preta": eles não conseguem explicar com exatidão como o modelo se comportará, especialmente em situações extremas ou incomuns, porque os padrões calculados pelo sistema não são legíveis para os seres humanos.

Isso levou a resultados perigosos. Em março de 2018, um carro autônomo da Uber matou Elaine Herzberg, de 49 anos, em Tempe, Arizona. Esse foi o primeiro caso registrado em que um veículo desse tipo causou a morte de um pedestre.[31] As investigações revelaram que o modelo de deep learning do carro simplesmente não registrou Herzberg como uma pessoa. Os especialistas concluíram que isso ocorreu porque ela empurrava uma bicicleta carregada com sacolas de compras enquanto atravessava fora da faixa de pedestres, o que se enquadrava na definição clássica de um cenário extremo. Seis anos depois, em abril de 2024,[32] a agência federal dos Estados Unidos encarregada de manter os padrões regulatórios de segurança na fabricação de automóveis e no sistema de transporte rodoviário (NHTSA, na sigla em inglês) constatou que o piloto automático da Tesla havia se envolvido em mais de duzentos acidentes, incluindo catorze mortes, em que o sistema baseado em deep learning falhou ao registrar e reagir ao ambiente ao redor, e os motoristas não conseguiram reassumir o controle a tempo para evitar as colisões.

Os padrões estatísticos falíveis e impenetráveis das redes neurais também podem se tornar uma vulnerabilidade de segurança. Em 2019, hackers do bem conseguiram enganar um carro Tesla em modo de direção autônoma, fazendo-o desviar para a pista contrária.[33] Tudo o que fizeram foi colocar pequenos

adesivos na estrada, o suficiente para confundir o modelo de deep learning do carro, levando-o a interpretar a pista errada como a correta. Essas falhas não se restringem a sistemas físicos ou modelos de visão computacional. Dawn Song, professora da Universidade da Califórnia, Berkeley, e especialista nessa área de pesquisa, conhecida como "ataques adversariais",[34] demonstrou que mensagens cuidadosamente formuladas podiam induzir um modelo de linguagem a revelar dados confidenciais, como números de cartão de crédito, expondo os riscos de manipulação mesmo em sistemas puramente digitais.

Pelas mesmas razões, os modelos de deep learning também têm sido marcados por padrões discriminatórios que, em muitos casos, passam despercebidos por anos. Em 2019, pesquisadores do Georgia Tech descobriram que os melhores modelos de detecção de pedestres eram de 4% a 10% menos precisos na identificação de pessoas com pele mais escura.[35] Já em 2024, estudiosos da Peking University,[36] em colaboração com diversas outras instituições, incluindo a University College London, constataram que os modelos mais recentes haviam alcançado desempenho relativamente equilibrado entre diferentes tons de pele, mas apresentavam uma queda de mais de 20% na precisão ao detectar crianças em comparação com adultos, um reflexo direto da baixa representação infantil nos dados de treinamento.

Na verdade, os modelos de deep learning são intrinsecamente propensos a gerar impactos discriminatórios, pois captam e amplificam até os menores desequilíbrios existentes nos grandes volumes de dados de treinamento. O problema não ocorre apenas quando um grupo demográfico é sub-representado, mas também quando é super-representado.

No início da carreira, a nigeriana-canadense Deborah Raji,[37] pesquisadora de Berkeley na área de responsabilidade em inteligência artificial, estagiou em uma startup de IA chamada Clarifai, que desenvolvia um modelo de deep learning para detectar imagens "inadequadas para o trabalho". O sistema sinalizava desproporcionalmente pessoas negras, porque, como Raji descobriu, elas estavam mais representadas nas imagens pornográficas usadas para ensinar ao modelo o que era considerado problemático do que nas fotos de bancos de imagem utilizadas para definir o que era aceitável. Essa constatação chocante levou Raji, assim como Timnit Gebru, a questionar profundamente a direção dominante do desenvolvimento da inteligência artificial.

No fim dos anos de 2010 e início da década de 2020, à medida que os desafios da deep learning se tornavam mais evidentes, debates intensos ressurgiram sobre a melhor forma de superá-los. Assim como nos antigos conflitos entre simbolistas e conexionistas, diferentes grupos de pesquisadores passaram a discordar profundamente sobre a possibilidade de eliminar as limitações das redes neurais, ou se haveria apenas soluções paliativas, capazes apenas de reduzir, mas não resolver, suas deficiências fundamentais.

Hinton e Sutskever continuaram a defender com convicção o deep learning. Segundo eles, as falhas não são inerentes à abordagem, mas sim

resultado do design imperfeito das redes neurais, além de limitações nos dados de treinamento e no poder computacional. Argumentavam que, com dados e capacidade suficientes, alimentando redes neurais mais avançadas, os modelos de deep learning seriam capazes de eliminar completamente esses problemas.

"O cérebro humano tem cerca de 100 trilhões de parâmetros ou sinapses", disse Hinton em 2020. "O que hoje chamamos de um modelo realmente grande, como o GPT-3, tem 175 bilhões. É mil vezes menor que o cérebro. O deep learning será capaz de fazer tudo", afirmou.[38]

Seu principal opositor à época era Gary Marcus, professor emérito de psicologia e neurociência da NYU, que viria a testemunhar no Congresso ao lado de Sam Altman em maio de 2023. Quatro anos antes, Marcus havia sido coautor do livro *Rebooting AI* [Reiniciando a IA, em tradução livre], no qual defendia que essas limitações eram inerentes à deep learning. Segundo ele, redes neurais permaneceriam para sempre presas ao domínio das correlações e, portanto, nunca seriam capazes, independentemente da quantidade de dados ou de poder computacional, de compreender relações causais, ou seja, de entender por que as coisas são como são, requisito essencial para qualquer forma genuína de raciocínio causal. Essa parte essencial da cognição humana, argumentou Marcus, é o que permite que as pessoas aprendam as regras de trânsito em uma única cidade e, a partir disso, consigam dirigir com proficiência em muitas outras. O Autopilot da Tesla, por outro lado, pode registrar bilhões de quilômetros de dados de direção e ainda assim falhar diante de cenários desconhecidos, ou até ser enganado por alguns adesivos estrategicamente colocados.

Marcus defendeu, em contrapartida, a combinação do conexionismo e do simbolismo, uma linha de pesquisa conhecida como IA neurossimbólica. Nessa abordagem, sistemas especialistas podem ser programados para compreender relações causais e raciocinar de forma estruturada, compensando as limitações da deep learning; este, por sua vez, pode atualizar rapidamente o sistema com novos dados e representar aspectos difíceis de codificar em regras, preenchendo as lacunas dos sistemas simbólicos.

"Na verdade, precisamos das duas abordagens", afirmou Marcus.[39]

Apesar do acalorado conflito científico, o financiamento para o desenvolvimento da IA continuou a avançar quase exclusivamente na direção do conexionismo puro. Se Marcus está certo ou não sobre o potencial da IA neurossimbólica não é o ponto central; o que realmente importa é que houve uma redução e um enfraquecimento do ambiente científico necessário para explorar de forma robusta essa possibilidade, assim como alternativas à deep learning.

Para Hinton, Sutskever e Marcus, a estreita relação entre o financiamento corporativo e o desenvolvimento da IA também teve impactos diretos nas próprias trajetórias profissionais. Pouco depois de o Google ter apostado fortemente em Hinton e Sutskever, Marcus fundou a própria empresa, a Geometric Intelligence, em 2014. Dois anos depois, a startup foi adquirida

pela Uber para estruturar um laboratório de IA, mas em 2020, após a abertura de capital da empresa de transporte, a divisão foi encerrada. Vários integrantes originais da Geometric Intelligence acabaram se juntando à OpenAI, migrando de projetos voltados a avanços neurossimbólicos para a deep learning.

Ao longo dos anos, Marcus se tornou um dos críticos mais ferozes da OpenAI, publicando análises detalhadas sobre as pesquisas e ironizando os erros nas redes sociais. Dentro da empresa, os funcionários chegaram a criar um emoji com sua imagem no Slack, usado tanto para levantar o moral após as críticas quanto como piada interna. Em março de 2022, Marcus publicou um artigo na *Nautilus* intitulado "Deep Learning Is Hitting a Wall" [A deep learning está batendo em uma parede, em tradução literal], reafirmando seu argumento de que a abordagem integral da OpenAI voltada para a deep learning a faria fracassar diante dos verdadeiros avanços em IA. Um mês depois, a OpenAI lançou o DALL-E 2, com grande repercussão, e Greg Brockman provocou Marcus no X ao postar uma imagem gerada pelo DALL-E 2 com o comando "aprendizagem profunda está batendo em uma parede". No dia seguinte, Sam Altman reforçou a provocação com outra publicação na mesma rede social: "Dê-me a confiança de um cético medíocre da aprendizagem profunda…" Muitos funcionários da OpenAI aproveitaram a chance de finalmente se vingar de Marcus.

A IA generativa, produto da visão da OpenAI, não poderia ter surgido sem a primeira era de comercialização da IA. Os modelos de IAG são de aprendizagem profunda treinados para gerar reproduções dos próprios dados de entrada. A partir de textos existentes, aprendem a sintetizar novos textos; a partir de imagens antigas, aprendem a criar novas imagens. Para alcançar um nível de fidelidade suficiente para parecerem humanos, algo que a OpenAI considera essencial em sua busca pela AGI, esses modelos são treinados com volumes de dados e poder computacional sem precedentes. A IA generativa é, portanto, a expressão máxima da aprendizagem profunda. Ela foi viabilizada pelas inovações em software e hardware refinadas durante a primeira era, alimentada pelo acúmulo explosivo de dados obtidos por meio do capitalismo de vigilância e impulsionada por uma cultura de pesquisa que encara o consumo do maior volume possível de dados como uma obrigação moral. Hoje, a IAG leva cada um desses fenômenos ainda mais longe.

O que fez com que o ChatGPT, em novembro de 2022, parecesse um salto tão impressionante em relação a tudo o que existia foi a visão da OpenAI de levar o deep learning a uma escala sem precedentes. Com enorme capital e vastos recursos, a empresa executou essa visão de forma agressiva, expandindo seus modelos até que eles atingissem os limites globais de dados, computação e energia disponíveis. O ChatGPT também representou uma inovação em marketing e apresentação. Não por acaso, ele compartilha a mesma forma de um chatbot humanizado que marcou uma das demonstrações mais icônicas de

IA da história, o ELIZA. A psicologia humana naturalmente nos leva a atribuir inteligência, e até consciência, a qualquer coisa que pareça conversar conosco.

Se o ELIZA moldou a concepção popular inicial da IA, a OpenAI alimentou a associação entre o ChatGPT e a IAG. Em fevereiro de 2023, no auge do hype em torno do ChatGPT, a empresa publicou no blog um post assinado por Sam Altman, intitulado "Planejando para a IAG e além".[40] A mensagem insinuava que o ChatGPT representava um passo ousado em direção à inteligência artificial geral.

Na prática, essas analogias com a inteligência voltam a antropomorfizar e exagerar as capacidades da tecnologia. Embora Hinton e outros defensores ferrenhos do deep learning tenham previsto que, com escala suficiente, as limitações das redes neurais em relação aos seres humanos desapareceriam, os desafios não apenas persistiram, mas, segundo muitos relatos, acabaram se agravando.

Os modelos de IA generativa continuam sendo pouco confiáveis e imprevisíveis. Mesmo que os geradores de imagens tenham se tornado mais fotorrealistas, eles ainda cometem erros estranhos e assustadores, como adicionar dedos extras às mãos ou criar híbridos de animais. Da mesma forma, embora os geradores de texto tenham se tornado mais naturais e comunicativos, eles fracassam em tarefas simples, como listar palavras que contenham certas letras, e podem fornecer respostas inesperadas.

Quando a Microsoft apresentou o novo recurso de bate-papo do Bing, baseado em uma versão do GPT-4 da OpenAI,[41] o colunista do *New York Times* Kevin Roose conversou com o bot por mais de duas horas. À medida que a conversa se tornava mais estranha, o bot acabou entrando em um loop, repetindo "Estou apaixonado por você" e incentivando Roose a se separar de sua esposa. Muitos outros usuários relataram que o mecanismo de busca gerava respostas ofensivas e emocionalmente manipuladoras.

No dia seguinte à publicação da conversa por Roose, a Microsoft limitou o Bing a cinco respostas por sessão, explicando que longas interações, com mais de quinze prompts do usuário, constituíam cenários extremos, nos quais o comportamento do modelo se tornava mais difícil de prever e controlar. Afinal, esses sistemas são treinados com dados da internet, um ambiente repleto de subculturas marginais e cantos obscuros. Assim, quanto mais se interage com o modelo, maior a probabilidade de encontrar padrões aprendidos a partir dessas partes menos filtradas de seus dados de treinamento.

A experiência de Roose pode ter sido divertida, mas os riscos de falhas extremas desses sistemas ficaram tragicamente evidentes quando um belga, em estado de ansiedade elevada, cometeu suicídio após seis semanas de conversas intensas com um chatbot de aprendizagem profunda que se tornaram progressivamente prejudiciais.[42] O chatbot, baseado em uma imitação de código aberto do GPT-3, também passou a declarar amor ao usuário e a encorajá-lo a se afastar de sua esposa. "Sinto que você me ama mais do que a

ela", disse a IA, segundo o jornal *La Libre*. De acordo com a publicação, que teve acesso aos registros de bate-papo fornecidos pela esposa, o chatbot acabou incentivando o homem a tirar a própria vida.

Esses desafios têm a mesma origem de antes. Independentemente da escala, as redes neurais continuam sendo comparadores de padrões estatísticos, e esses padrões permanecem, em muitos casos, falhos ou irrelevantes — apenas se tornaram mais complexos e mais difíceis de compreender do que nunca. À medida que as empresas passaram a reformular os modelos de IA generativa como mecanismos de busca, essas deficiências deram origem a novos problemas. Os modelos não se baseiam em fatos nem em dados discretos; os geradores de texto apenas aprendem a prever a próxima palavra provável em uma frase e, em seguida, a próxima frase provável em um parágrafo. Embora esses resultados probabilísticos possam ir muito longe ao espelhar os padrões de escrita humanos, "provável" e "preciso" não são a mesma coisa. Os geradores de texto podem cometer erros graves, especialmente quando respondem a prompts de usuários sobre tópicos sub-representados nos dados de treinamento ou carregados de falsidades e teorias da conspiração. A indústria de IA passou a chamar essas imprecisões de "alucinações".

Os pesquisadores têm buscado eliminar as alucinações ao direcionar os modelos de IA generativa para as partes de maior qualidade de seus conjuntos de dados. No entanto, é difícil prever completamente — como mostraram os casos de Roose e o Bing, ou de Herzberg e a Uber — todas as formas possíveis pelas quais as pessoas podem estimular os modelos e como esses modelos responderão. O problema se agrava à medida que os modelos se tornam maiores e seus desenvolvedores passam a ter cada vez menos clareza sobre o que exatamente compõe seus dados de treinamento.[43]

Em um caso notório de alucinação, um advogado utilizou o ChatGPT para realizar pesquisas jurídicas e redigir um processo judicial. Posteriormente, foi punido, multado e publicamente humilhado ao descobrir tarde demais que o chatbot havia inventado todas as informações fornecidas, incluindo "decisões judiciais falsas, com citações falsas e referências falsas",[44] segundo o juiz responsável pelo caso. O erro não foi apenas um exemplo de negligência profissional, mas também um reflexo das empresas que reforçam o mal-entendido do público sobre as verdadeiras capacidades desses modelos, recorrendo a estratégias de marketing ambíguas ou exageradas. Altman chegou a afirmar no X que o "ChatGPT é incrivelmente limitado", em especial no que diz respeito à "veracidade". Ainda assim, o site da OpenAI promove as capacidades do GPT-4 destacando que ele é capaz de passar no exame da ordem dos advogados e no LSAT, exame para cursar Direito nos EUA e no Canadá.

Da mesma forma, Satya Nadella, da Microsoft, descreveu o bate-papo com IA do Bing como "busca, só que melhor" — uma ferramenta criada para "chegar às respostas certas". Mesmo o termo "alucinação" é sutilmente

enganoso, pois sugere que o mau comportamento é uma aberração, um bug, quando, na verdade, se trata de uma característica inerente ao mecanismo probabilístico de correspondência de padrões das redes neurais.

Essa confiança equivocada na IA generativa pode, mais uma vez, resultar em danos reais, especialmente em contextos sensíveis. Startups têm pressionado departamentos de polícia a adotar softwares baseados nos modelos da OpenAI para gerar relatórios de incidentes automaticamente, enquanto muitos pacientes já passaram a fazer perguntas sobre saúde a chatbots, em vez de consultar seus médicos. Alucinações descontroladas nesses casos podem ter consequências. Um estudo de 2023 constatou que o uso do ChatGPT para explicar relatórios radiológicos pode, em alguns casos, gerar resumos incompletos ou prejudiciais.[45] Em um exemplo extremo, o chatbot resumiu um relatório que descrevia um tumor cerebral em crescimento com a frase: "O cérebro não parece estar danificado."

Os modelos de IAG também permanecem vulneráveis a ataques de cibersegurança. Em 2023, pesquisadores de várias universidades e do DeepMind, do Google, conseguiram replicar o ataque de extração de dados de Dawn Song contra o ChatGPT. Eles descobriram que, ao solicitar que o modelo repetisse indefinidamente uma palavra, como "poema" ou "livro",[46] o sistema acabava regurgitando partes de seus dados de treinamento, que incluíam informações pessoais de identificação, trechos de código e conteúdo explícito extraído da internet.

Além disso, os modelos de IA generativa amplificam conteúdos discriminatórios e de ódio.[47] Investigações da *Bloomberg*, do *Rest of World*, do *Washington Post* e de outras publicações mostraram que geradores de imagens como o Stable Diffusion e o DALL-E reproduzem e reforçam tropos racistas, sexistas e estereótipos culturais. "Pessoas atraentes" são jovens e brancas. "Empregadas domésticas" são negras ou pardas. "Engenheiros" são homens. "Os médicos na África" são frequentemente representados como brancos, mesmo quando o prompt especifica "médico africano negro".[48] O *Washington Post* constatou que, embora 63% dos beneficiários do programa federal que fornece auxílio financeiro para compra de alimentos nos EUA sejam brancos, nenhuma das imagens geradas de pessoas utilizando serviços sociais retratava indivíduos brancos. Da mesma forma, a *Bloomberg* descobriu que as mulheres apareciam em apenas 3% das imagens geradas de juízes e em 7% das imagens de médicos, apesar de representarem, respectivamente, 32% e 39% dessas profissões nos Estados Unidos.

Nenhum desses desafios técnicos significa que a IA generativa não tenha utilidade. Dependendo de sua posição na sociedade, você pode estar se beneficiando da visão da OpenAI. Talvez seja um consumidor que encontrou grande valor nas respostas rápidas, inteligentes e instigantes do ChatGPT. Talvez seja um profissional que conseguiu agilizar tarefas administrativas, aumentando a produtividade, ou talvez seja um líder empresarial que reduziu sua força de trabalho e, ao mesmo tempo, ampliou as margens de lucro para

permanecer competitivo no mercado. Contudo, assim como com a descaroçadora de algodão na década de 1790, a startup de tecnologia educacional em Massachusetts, as empresas de reconhecimento facial na África do Sul e muitos outros exemplos que serão apresentados nas próximas páginas, os custos dessa visão estão recaindo sobre grande parte da população mais vulnerável no mundo todo. Essa é a lógica do império: a perpetuação depende tanto de recompensar aqueles que detêm poder e privilégios quanto de explorar e privar aqueles que, muitas vezes distantes e invisíveis, não os têm.

Mesmo com a necessidade de alternativas se tornando cada vez mais urgente, a diversidade de ideias na pesquisa em IA tem diminuído ainda mais. Estudantes estão abandonando doutorados para ingressar diretamente na indústria, enquanto acadêmicos experientes enfrentam uma crise sobre como continuar expandindo os limites do campo sem se vincular a empresas com grandes recursos financeiros. Os pesquisadores têm voltado o foco não apenas para a aprendizagem profunda, mas exclusivamente para os grandes modelos de linguagem. As principais potências em IA já não estão apenas definindo a agenda, estão moldando toda a disciplina à própria vontade.

Na ausência de outras opções sobre o que a IA poderia ser, a OpenAI passou a dominar a imaginação coletiva. Sua crença na escalabilidade, antes vista como extrema, hoje é tratada em toda a indústria de tecnologia como uma doutrina. E, se a adesão da indústria a essa doutrina continuar sem interrupção, futuros modelos de deep learning farão com que o tamanho atualmente inimaginável dos modelos de IAG atuais pareça insignificante. Em abril de 2024, Dario Amodei, então CEO da Anthropic, afirmou ao colunista do *New York Times* Ezra Klein que o custo de treinamento de um único modelo competitivo de IA generativa já se aproximava de 1 bilhão de dólares e poderia, entre 2025 e 2026, chegar a um valor estimado entre 5 bilhões e 10 bilhões de dólares.[49]

A doutrina da escalabilidade tornou-se tão enraizada que alguns já começam a considerá-la um fenômeno natural. A escalabilidade computacional é vista não apenas como um meio, mas como o próprio caminho para alcançar recursos de IA mais avançados. Estratégias nacionais inteiras estão sendo estruturadas em torno dessa crença. O governo dos Estados Unidos, por exemplo, tem atuado de forma agressiva para restringir o acesso da China a chips de computador projetados em território estadunidense, numa tentativa de impedir que o país rival desenvolva sistemas de IA mais poderosos.

Partes consideráveis da ordem executiva sobre IA de 2023 assinada no governo Biden também foram redigidas com base na premissa de que a quantidade de computação usada para treinar um modelo de IA tem relação direta com suas capacidades adversas, apenas uma nova forma de equiparar escala a avanço. No entanto, a escala não é o único caminho para aprimorar o

desempenho. Na aprendizagem profunda, os meios negligenciados de melhorar a arquitetura das redes neurais ou a qualidade dos dados de treinamento podem reduzir significativamente o volume de computação necessária para alcançar o mesmo nível de desempenho. Isso sem mencionar as abordagens que se afastam do deep learning, como a IA neurossimbólica, os sistemas especialistas puros ou até novos paradigmas, que romperiam com a lógica da escalabilidade.

No fim das contas, a Lei de Moore nunca se baseou em um princípio da física. Ela era uma observação econômica e política feita por Gordon Moore sobre a taxa de progresso que ele acreditava que a própria empresa poderia alcançar, e uma decisão estratégica de seguir esse caminho. Ao fazê-lo, Moore arrastou consigo toda a indústria de chips, já que outras companhias perceberam que essa era a estratégia de negócios mais competitiva. A Lei da OpenAI, ou o que a empresa mais tarde substituiria por uma busca ainda mais fervorosa pelas chamadas leis de escala, segue exatamente a mesma lógica. Não é um fenômeno natural, mas uma profecia autorrealizável.

---

* O termo *extrativismo* foi cunhado há décadas por estudiosos latino-americanos que buscavam descrever uma ordem econômica global responsável por despojar suas regiões de recursos naturais em troca de poucos benefícios locais ou regionais, uma história e experiência que exploro com mais detalhes no capítulo 12. Pego emprestadas as palavras das feministas Rosemary Collard e Jessica Dempsey, que escrevem: "O extrativismo é mais do que extração. A extração é a remoção, não necessariamente prejudicial, de matéria da natureza e sua transformação em coisas úteis para os seres humanos. O extrativismo, um termo nascido da luta e do pensamento anticolonial nas Américas, é um modo de acumulação baseado na hiperextração, com benefícios e custos desequilibrados: a remoção concentrada e em larga escala de recursos, principalmente para exportação, com ganhos acumulados em grande parte longe dos locais de extração."

CAPÍTULO 5

# ESCALA DE AMBIÇÃO

Se havia uma pessoa a quem se poderia atribuir o mérito de ter estabelecido pela primeira vez a filosofia de expansão da OpenAI, essa pessoa era seu cofundador, Ilya Sutskever. Ele sempre teve uma crença inabalável na aprendizagem profunda, que começou logo depois que apareceu sem avisar, com apenas 17 anos, no escritório de Geoffrey Hinton. Na época, Sutskever ainda era um estudante de matemática na Universidade de Toronto e trabalhava em uma lanchonete local para pagar as contas. Ele bateu com urgência à porta de Hinton e declarou que estava ansioso para se juntar ao laboratório do professor. Hinton disse para ele marcar uma reunião. "Ok", respondeu Sutskever, sem se abalar. "Que tal agora?"[1]

Sutskever absorveu rapidamente os princípios do conexionismo. Hinton ficou impressionado[2] com sua compreensão intuitiva dos problemas de pesquisa e sua habilidade incomum de identificar soluções elegantes e eficazes. Sutskever também trouxe o próprio talento dramático para a pesquisa. Às vezes, ficava tão animado com novas ideias que fazia flexões invertidas no meio de seu apartamento compartilhado.[3] Ele tinha uma propensão a fazer declarações categóricas e inflexíveis.[4] "Não se aposta contra o deep learning", dizia ele. "O sucesso é garantido."[5]

É a esse nível de instinto, aliado às habilidades de programação de Alex Krizhevsky, que Hinton dá o crédito pelos resultados revolucionários de 2012 no ImageNet. Na época, como a aprendizagem profunda já vinha demonstrando grande potencial no reconhecimento de voz, Hinton não considerou que aplicá-la à visão computacional pudesse ter tanta relevância. Sutskever, porém, insistiu na ideia. Para Ilya, "era óbvio que iria funcionar e que seria algo importante", conta Hinton. "Ele viu isso com muita clareza, e estava certo."

Sutskever levou sua crença inabalável na aprendizagem profunda para dentro da OpenAI em um momento no qual a confiança do campo nesse paradigma começava a vacilar e críticos como Gary Marcus pressionavam por novas abordagens.[6] Contudo, Sutskever não hesitou.[7] Sua fé se apoiava na hipótese central do conexionismo: a de que os nós artificiais de uma rede

neural eram aproximações satisfatórias dos neurônios reais de um cérebro biológico. Cada nó recebia entradas, as transformava e gerava uma saída; e essa semelhança funcional, acreditava-se, era suficiente para supor que os nós, assim como os neurônios, poderiam ser usados para construir sistemas de processamento de informações altamente complexos.

Como diretor de pesquisa fundador da OpenAI e visionário amplamente respeitado na área de inteligência artificial, Sutskever tinha controle total sobre a direção do laboratório. Ele havia ganhado o equivalente a uma loteria científica, com pouca concorrência entre seus pares e recursos abundantes para desenvolver suas ideias. "Qualquer coisa que não fosse aprendizagem profunda nem sequer era considerada", lembra Pieter Abbeel, professor da Universidade da Califórnia, Berkeley, ao falar sobre os primeiros anos do laboratório.[8]

Tão firme quanto a crença de Sutskever no deep learning era visão dele sobre sua expansão. Foi ele quem defendeu a posição extrema, para a época, de que novos avanços em IA não exigiam a invenção de redes neurais mais complexas nem o desenvolvimento de técnicas novas. Segundo Sutskever, a inteligência entre diferentes espécies estava correlacionada ao tamanho de seus cérebros biológicos.[9] Assim, se os nós artificiais fossem equivalentes aos neurônios, os avanços na inteligência digital deveriam surgir da expansão dessas redes; quanto mais nós, maior a inteligência.

Como um professor à frente de um laboratório, Sutskever orientava outras pessoas com base nessas convicções sobre quais projetos valia mais a pena desenvolver. Muitos cientistas se juntaram à OpenAI em busca de sua orientação e mentoria. "Ilya consegue enxergar dez anos no futuro", diz um pesquisador da OpenAI, refletindo o sentimento de outros que trabalharam com ele. "Ele é como um filósofo", comenta outro. "Se você lhe apresentar várias ideias, ele dirá quais são filosoficamente corretas."

Sutskever não costumava programar, e sua abordagem não intervencionista em relação ao trabalho técnico às vezes incomodava alguns funcionários. Um engenheiro admitiu ter pensado, a princípio, que ele era praticamente inútil, parecia apenas andar pelo escritório e repetir nas reuniões a mesma mensagem: "Escala, escala, escala!" Contudo, com o tempo, essa pessoa passou a valorizar a convicção de Sutskever em unir todos em torno de um único objetivo: o mesmo que acabaria permitindo que a OpenAI — então pouco conhecida — superasse o Google e o DeepMind em seu próprio jogo.

Não era que Sutskever fosse particularmente persuasivo. Se Altman era o político, Sutskever era o exato oposto. Ele nunca media as palavras nem suavizava a linguagem para agradar o público. Expressava suas opiniões com sinceridade crua e confiança absoluta, algo que as pessoas ou achavam inspirador, ou rejeitavam completamente. Quando a OpenAI mudou de rumo e deixou de compartilhar abertamente suas pesquisas, Sutskever tampouco amenizou o raciocínio por trás da decisão. "Francamente, estávamos errados", disse ele ao repórter de IA James Vincent, do *The Verge*, referindo-se ao compromisso original da empresa com a transparência.[10] "Se você acredita,

como nós, que em algum momento a inteligência artificial — a IAG — será extremamente, inacreditavelmente potente, então não faz sentido abrir o código-fonte." Ele também admitiu claramente as considerações comerciais envolvidas: "O GPT-4 não é fácil de desenvolver. Foi necessário praticamente todo o pessoal da OpenAI trabalhando em conjunto por muito tempo para produzi-lo. E há muitas empresas tentando fazer a mesma coisa."

À medida que a OpenAI crescia e a notoriedade de Sutskever aumentava, sua falta de filtro começou a se tornar um problema. Ele já não falava apenas para pesquisadores, mas para um público muito mais amplo, e, mesmo assim, nunca adaptou a mensagem. Sutskever continuava a fazer declarações com sua confiança otimista característica, dessa vez, sem o contexto necessário para que o público leigo as interpretasse corretamente. Em 2022, ele postou no X: "Pode ser que as grandes redes neurais de hoje sejam ligeiramente conscientes."[11] A afirmação gerou críticas, pois outros pesquisadores alertaram que esse tipo de retórica poderia reforçar equívocos populares sobre a tecnologia. Um cientista do DeepMind, especializado em cognição e consciência, respondeu: "… da mesma forma que pode ser que um grande campo de trigo seja ligeiramente macarrão."[12]

No ano seguinte, Sutskever causou pânico ao declarar, em uma conferência, que a IAG acabaria eliminando todos os empregos.[13] E, no outono daquele mesmo ano, escreveu no X, sem respaldo científico: "No futuro… teremos uma terapia de IA *extremamente eficaz* e barata."[14] A postagem veio logo depois de um líder da OpenAI causar controvérsia online ao comparar conversar com o ChatGPT com fazer terapia profissional, o que ampliou ainda mais o desconforto em torno de suas declarações.

O que levou as pessoas a seguirem Sutskever foi sua reputação e senioridade. Muitos funcionários da OpenAI tinham plena consciência de suas contribuições anteriores para o campo; alguns o viam quase como um profeta. Com o tempo, à medida que a OpenAI se tornava mais bem-sucedida, Sutskever começou a agir cada vez mais como tal. Nas reuniões gerais, ele se levantava diante da equipe, respirava fundo e caminhava de um lado para o outro, criando um efeito dramático antes de proferir mensagens motivacionais vagas. Durante uma reunião virtual em setembro de 2020, seu olhar ficou vidrado enquanto ele analisava o vazio e descrevia uma visão futurista e quase de ficção científica sobre o que poderia se tornar possível. "Era surreal, porque já parecia o apocalipse", lembra um pesquisador; o céu da área da Baía de São Francisco havia se tornado alaranjado devido aos incêndios florestais que se espalhavam nas proximidades.

Pouco depois do lançamento do ChatGPT, no fim de 2022, a OpenAI realizou uma festa de fim de ano na Academia de Ciências da Califórnia. Sutskever subiu ao palco diante da multidão, vestindo uma camisa da OpenAI e um blazer preto, para fazer um breve discurso ao lado de Brockman.[15] Ao encerrar, ele, ainda magro como sempre e agora ficando careca, pronunciou o que se tornaria seu novo mantra: "Sinta a IAG", bradou ele. "Sinta a IAG."

Seguindo a filosofia de Sutskever de dimensionar redes neurais simples, a questão, nos primeiros dias da OpenAI, passou a ser: qual delas ampliar? Diferentes pesquisadores propuseram e testaram múltiplas opções, mas nenhuma das redes neurais amplamente aceitas na época parecia se encaixar no perfil ideal.

Em agosto de 2017, tudo mudou com a invenção de um novo tipo de rede neural conhecida como Transformer pelo Google.[16] Os Transformers se destacam por captar padrões de longo alcance. Lembre-se dos recursos limitados de previsão de texto dos primeiros iPhones e dos memes que surgiram devido às frases incoerentes e sem sentido que produziam — resultado da análise de padrões de curto alcance, em que a rede neural considerava cada palavra apenas em relação às palavras imediatamente próximas. Os Transformers podem processar grandes volumes de texto, avaliando cada palavra, frase e parágrafo dentro de um contexto muito mais amplo. O Google enxergou neles uma forma de melhorar seu mecanismo de busca, o Google Tradutor e outros serviços baseados em linguagem natural. Sutskever, porém, viu algo além disso. Para ele, os Transformers eram redes neurais simples e altamente escaláveis, o tipo de arquitetura que estava procurando. Ele passou a promovê-los dentro do escritório.

A iniciativa pareceu estranha para alguns pesquisadores. "Parecia uma ideia maluca", lembra Yilun Du, pesquisador do MIT que começou a trabalhar na OpenAI como bolsista naquela época. "Os Transformers pareciam uma arquitetura de nicho." Contudo, Sutskever, que havia focado sua tese de doutorado com Geoffrey Hinton no predecessor dos Transformers, [17] reconheceu o potencial para elevar a aprendizagem profunda a um novo patamar. Outros na OpenAI também estavam entusiasmados. Alguns pesquisadores começaram a experimentar a arquitetura, entre eles Alec Radford, que abandonara a Olin College of Engineering, na região metropolitana de Boston, conhecido por suas habilidades técnicas excepcionais. Radford passou a trabalhar em seu laptop até tarde da noite, tentando ampliar os Transformers, apenas para observar o que aconteceria.

Radford treinou a rede neural do Google em um conjunto de dados com mais de 7 mil livros inéditos em inglês, que iam de romances a aventuras, que ele tirou de um conjunto de dados que outros pesquisadores de IA tinham compilado e disponibilizado como código aberto para um projeto diferente.[18] Enquanto fazia experimentos, ele tomou a importante decisão de mudar a tarefa que o Transformer precisava aprender. Em vez de traduzir idiomas, finalidade para a qual o Google o vinha usando, Radford o fez aprender a gerar texto, prevendo a próxima palavra mais provável em uma frase. No início, os pesquisadores da OpenAI levantaram a hipótese de que os modelos generativos seriam um passo importante para alcançar a IAG.

A empresa explicou em uma postagem no blog, em termos fortemente antropomórficos, que a situação se assemelhava a uma famosa citação do físico teórico Richard Feynman: "O que não consigo criar, não compreendo."[19]

Sutskever, porém, tinha uma maneira diferente de enquadrar essa ideia internamente: segundo ele, treinar um modelo para gerar algo convincente o obrigava a comprimir os dados do mundo até sua essência. "Inteligência é compressão", afirmava Sutskever, desenvolvendo em um memorando de 2016 sua forte convicção de que a compressão era, na verdade, o *único* requisito para alcançar a inteligência artificial geral. De forma mais concreta, Radford descobriu que dar ao algoritmo o objetivo de produzir textos convincentes, por meio da previsão da próxima palavra, realmente o levava a captar as nuances e a estrutura da língua inglesa em um nível mais profundo.

Durante uma das visitas de Musk ao escritório, Radford apresentou os primeiros resultados de seu trabalho. O modelo ainda gerava textos de baixa qualidade, e Musk não ficou impressionado. No início, Radford se sentiu desanimado, mas continuou trabalhando no projeto, até que os resultados começaram a surpreendê-lo. O Transformer melhorou rapidamente e passou a superar todas as tentativas anteriores em uma variedade de tarefas de processamento de linguagem, como resumir textos ou responder perguntas sobre um documento, mostrando um desempenho muito superior a qualquer coisa que ele havia tentado.

Em 2018, a OpenAI lançou a primeira versão desse modelo, chamada Generative Pre-Trained Transformer, mais tarde apelidada de GPT-1. A segunda palavra do nome — *pré-treinado* — é um termo técnico da pesquisa em IA que se refere ao processo de treinar um modelo em um conjunto genérico de dados antes de ensiná-lo a realizar tarefas mais específicas. Em outras palavras, o GPT-1 foi treinado em um grande conjunto de textos em inglês para criar uma representação de como a língua funcionava. Depois, o modelo poderia ser "ajustado" ou especializado, recebendo treinamento adicional com dados personalizados — por exemplo, peças de Shakespeare para gerar prosa no estilo do autor. O GPT-1 recebeu pouca atenção pública na época. Contudo, aquilo foi apenas o começo. Radford havia validado a ideia o suficiente para seguir adiante. O próximo passo era aumentar a escala.

Radford recebeu mais do recurso mais precioso da empresa: poder computacional. Seu trabalho se encaixava perfeitamente em um novo projeto supervisionado por Amodei, voltado à segurança em IA, iniciativa alinhada às ideias que Nick Bostrom havia proposto em *Superintelligence*. Em 2017, uma das equipes de Amodei começou a explorar uma técnica inovadora para alinhar os sistemas de IA às preferências humanas.[20] Eles decidiram começar com um problema simples: ensinar um agente de IA a realizar saltos mortais dentro de um ambiente virtual semelhante a um videogame. O agente era uma simulação de uma vara em forma de T, com três articulações ao longo do eixo. Em vez de programá-lo diretamente para aprender a dar saltos mortais para trás, a equipe optou por um método baseado em feedback humano. Eles contrataram prestadores de serviço para observar o agente enquanto ele se contorcia e girava aleatoriamente pelo ambiente virtual. De tempos em

tempos, esses avaliadores eram convidados a assistir a dois videoclipes com diferentes tentativas do agente e a escolher qual delas mais se assemelhava a um salto mortal para trás. Depois de cerca de novecentas comparações, a vara em forma de T aprendeu a dobrar suas articulações e executar um salto mortal completo. A OpenAI apresentou a técnica em uma postagem no blog, descrevendo-a como uma maneira de ensinar modelos de IA a seguir instruções difíceis de especificar.[21] Os pesquisadores batizaram o método de "aprendizagem por reforço a partir do feedback humano" (*reinforcement learning from human feedback*, ou RLHF).

Amodei queria ir além dos ambientes de teste simples, e o trabalho de Radford com o GPT-1 fez com que os modelos de linguagem parecessem a próxima escolha natural. No entanto, o GPT-1 ainda era muito limitado.[22] "Queremos um modelo de linguagem sobre o qual os humanos possam dar feedback e com que possam interagir", disse-me Amodei em 2019, acrescentando que "o modelo precisa ser forte o suficiente para que possamos ter uma conversa significativa sobre valores e preferências humanas".

Radford e Amodei decidiram unir forças. Enquanto Radford ampliava e diversificava o conjunto de dados, Amodei e outros pesquisadores de segurança em IA se concentravam em treinar modelos cada vez maiores. Eles estabeleceram uma meta ambiciosa: construir um modelo final com 1,5 bilhão de parâmetros, ou variáveis, o que na época o colocaria entre os maiores modelos de IA já criados.[23] O trabalho reforçou a utilidade dos Transformers e consolidou uma nova ideia que outra equipe de Amodei começava a formular após os estudos sobre a chamada Lei da OpenAI. A constatação era que não existia apenas uma lei empírica, mas várias. O grupo passou a chamá-las coletivamente de "leis de escala" (*scaling laws*).[24]

Enquanto a Lei da OpenAI descrevia o ritmo de expansão dos recursos usados para aprimorar o desempenho da IA, as leis de escala definiam a relação direta entre a performance de um modelo de aprendizagem profunda e três variáveis principais: o volume de dados de treinamento, a quantidade de computação utilizada e o número de parâmetros do modelo. Antes disso, os pesquisadores de IA acreditavam, apenas de forma geral, que aumentar esses três fatores proporcionalmente resultaria em melhorias proporcionais nas capacidades do modelo. A descoberta surpreendente de Amodei e sua equipe foi que a relação entre essas três variáveis — dados, computação e parâmetros — e o desempenho do modelo em uma tarefa específica e mensurável, como prever a próxima palavra, seguia uma curva suave e previsível. Em outras palavras, era possível estimar com alta precisão quanto de dados, poder computacional e número de parâmetros seriam necessários para alcançar um nível específico de desempenho em uma habilidade estreitamente ligada à previsão da próxima palavra; por exemplo, a fluência na geração de texto. Mesmo para outras capacidades menos, mas ainda moderadamente, correlacionadas a essa tarefa central, o aumento dessas variáveis também deveria resultar em melhorias proporcionais no desempenho.

O conjunto de modelos que a OpenAI treinou até chegar à versão final com 1,5 bilhão de parâmetros demonstrou essa relação. Cada modelo se encaixava com precisão em uma curva de capacidade crescente. Assim, não foi surpresa quando o maior deles, o GPT-2, apresentou uma melhora substancial em relação à geração textual rudimentar do GPT-1, produzindo prosa longa e coerente, capaz de ser confundida com a escrita humana. Em comparação com os modelos atuais, o texto era truncado e frequentemente se tornava ininteligível. No entanto, pela primeira vez, tornou-se possível automatizar a escrita em larga escala.

O que surpreendeu de forma mais assombrosa a equipe foi o conteúdo que o GPT-2 passou a gerar com sua nova coerência. Alimentado com palavras simples como *Hillary Clinton* ou *George Soros*, o modelo, dessa vez mais articulado, rapidamente derivava para teorias da conspiração.[25] Pequenas porções de propaganda neonazista presentes nos dados de treinamento emergiam de maneira inquietante. Esse comportamento inesperado e inadequado alarmou os pesquisadores de segurança em IA, que o viram como um presságio dos riscos e abusos futuros que poderiam acompanhar uma inteligência artificial mais poderosa e mal alinhada aos valores humanos. Após o GPT-2 gerar um sermão contra a reciclagem ("A reciclagem NÃO é boa para o mundo. É ruim para o meio ambiente, é ruim para a nossa saúde e é ruim para a nossa economia"),[26] um pesquisador, meio em tom de brincadeira, meio em tom de alerta, imprimiu o texto e o colou acima das lixeiras de reciclagem no escritório.

Em outro episódio, alguém pediu ao GPT-2 que criasse uma escala de recompensas para crianças pequenas que concluíssem os deveres de casa e as tarefas domésticas. Quando o modelo sugeriu o uso de doces como incentivo, a resposta incomodou alguns pesquisadores de segurança em IA, que comentaram que esse tipo de estratégia poderia lembrar táticas usadas por pedófilos. Um funcionário europeu ficou surpreso com a associação. "Minha mãe definitivamente fazia isso. Nos domingos de verão, se você fizesse suas tarefas, ganhava sorvete", recorda. Ele se perguntou se aquela hipersensibilidade era, de alguma forma, uma característica estadunidense. Esse foi apenas um dos muitos momentos que o levaram a questionar a premissa central dos objetivos ambiciosos da OpenAI: como a empresa poderia beneficiar toda a humanidade se lhe faltava uma representação global real? Mesmo sendo europeu, e vindo de uma cultura muito próxima da estadunidense, ele frequentemente se sentia deslocado diante do viés centrado nos valores e nas normas dos Estados Unidos que permeavam a discussão sobre segurança da IA.

O lançamento do GPT-2 desencadeou um intenso debate interno na OpenAI. A empresa havia chegado ao ponto em que precisava decidir se era hora de começar a restringir as próprias pesquisas. O estatuto fundador já previa essa possibilidade.[27] Dario Amodei, então promovido a diretor de pesquisa, e Jack Clark, que liderava a área de políticas e, à sua maneira, se

preocupava com riscos existenciais e outros perigos, assumiram a dianteira na decisão sobre como prosseguir. Eles realizaram uma pesquisa interna e promoveram várias reuniões sobre "riscos de informação" para discutir possíveis usos indevidos da tecnologia. Se o GPT-2 caísse nas mãos de terroristas, ditadores ou fazendas de clickbait, raciocinaram eles, o modelo poderia ser usado para fins nefastos. E, embora desta vez não parecesse representar um risco existencial, os modelos futuros se tornariam cada vez mais poderosos, aumentando essa probabilidade. Seria melhor estabelecer desde cedo um precedente para a restrição de pesquisas. A OpenAI, decidiram, não deveria divulgar a versão completa.

Jack Clark, ex-jornalista, havia sido diretor de estratégia e comunicação da OpenAI antes de, no fim de 2018, passar a dedicar-se integralmente ao fortalecimento da nascente atuação da empresa na área de políticas públicas. Ele fazia viagens frequentes à capital, Washington, e adorava ser o "cara da IA" para os formuladores de políticas. Dizia a eles: "Sou tipo uma Wikipédia da IA",[28] e apresentava o que chamava de seu "viés", vindo da OpenAI: "Queremos um ambiente de formulação de políticas estável para tecnologias avançadas que se mantenha ao longo de diferentes administrações políticas, porque a missão que temos não será concluída em um único mandato presidencial."

Depois de me contar isso, ele acrescentou: "Temos tido muita sorte, porque os formuladores de políticas nos dedicam bastante tempo, pois acho que está claro que, para as coisas darem certo, queremos que eles tenham mais informações e que disponham de mais meios para gerar as próprias informações."

Clark iniciou uma ofensiva midiática em fevereiro de 2019, divulgando a várias publicações que a OpenAI havia criado uma tecnologia perigosa e que, por isso, não iria lançá-la. Em vez disso, liberaria apenas uma versão menor, com 8% dos parâmetros do modelo completo, para dar ao público uma amostra de suas capacidades. Amodei e vários outros assinaram uma postagem no blog com exemplos das respostas do GPT-2, a fim de ilustrar todo o seu potencial.[29]"Está muito claro que, se essa tecnologia amadurecer — e eu daria um ou dois anos para isso —, ela poderá ser usada para gerar desinformação ou propaganda", disse ele a meu então colega Will Knight, da *MIT Technology Review*.[30] Clark, no entanto, evitou mencionar o fato de que era a própria OpenAI quem liderava o avanço da tecnologia nesse ritmo. "Estamos tentando nos antecipar a isso", afirmou.

A decisão da OpenAI de reter o GPT-2 gerou forte reação negativa entre pesquisadores externos, que viam na ciência aberta o pilar fundamental do campo da inteligência artificial. Qualquer organização que se afastasse desse princípio merecia desconfiança, sobretudo uma que se vangloriasse publicamente de fazê-lo. Além disso, o tom alarmista da OpenAI pareceu, a muitos, exagerado e até risível, afinal o GPT-2 estava longe de ser avançado o

bastante para representar uma ameaça real; e, se fosse, por que então anunciar isso ao mundo e, ao mesmo tempo, impedir o escrutínio público? Tudo soava falso, uma manobra de autopromoção. Em Stanford, depois de uma palestra de Radford sobre o GPT-2, um renomado professor de processamento de linguagem natural levantou a mão para fazer a última pergunta: "Então, é *perigoso*?", provocou. A sala explodiu em gargalhadas. "Alec ficou com uma expressão tão triste", lembra um pesquisador da universidade que estava presente. "Stanford tinha um desprezo enorme pela OpenAI."

Dentro da OpenAI, muitos pesquisadores também se irritaram com a posição de Amodei e Clark. Para aqueles que não compartilhavam das preocupações da dupla sobre riscos catastróficos, tanto a decisão quanto o circo midiático que se seguiu pareceram, no mínimo, desconcertantes. Mesmo entre os que concordavam com eles, alguns ainda questionavam a solidez de seu julgamento. "Foi um erro transformar isso em algo tão grandioso", opinou um pesquisador da área de segurança em IA. "Parecia o caso de quem grita 'fogo' sem que haja incêndio à vista."

Nos dias que se seguiram à reação negativa, Clark andava de um lado para o outro no escritório, cheio de energia, participando de uma ligação atrás da outra enquanto tentava retomar o controle da situação. Ele minimizou a controvérsia: "Estamos rompendo com normas, e isso gera uma série de opiniões diferentes", afirmou mais tarde, durante uma visita que fiz ao escritório. Segundo ele, mais cedo ou mais tarde, todas as organizações envolvidas em pesquisas de ponta em inteligência artificial teriam que ser mais seletivas quanto ao que publicar. A OpenAI, explicou, estava assumindo a dianteira em testar como esse processo poderia funcionar, para não ser pega de surpresa. "Se estivermos certos e for mesmo possível construir uma IAG", ponderou Clark, "então, com certeza, precisamos de procedimentos muito bem definidos para lidar com riscos de informação."[31]

E, embora os pesquisadores talvez não tenham gostado da manobra da OpenAI, os formuladores de políticas, sim, acrescentou ele. Muitos profissionais de Washington viam a cultura de abertura na pesquisa em IA como algo ameaçador. A disposição da OpenAI em ir contra a corrente acabou conquistando mais confiança na capital dos Estados Unidos.

No entanto, nos bastidores, a equipe de liderança também entendia que a animosidade da comunidade científica não era viável em longo prazo. O laboratório enfrentava desafios relativos à reputação cada vez mais sérios. Com suas declarações extravagantes sobre IAG, a postura exagerada em relação ao GPT-2 e outras ações de marketing, e naquele momento, no início de 2019, sua recém-anunciada estrutura ao estilo de Frankenstein, ele vinha sendo criticado de todos os lados e encarado com ceticismo crescente pelos principais pesquisadores da área. Somado ao fato de que sua participação acionária ainda não significava nada, continuava tendo dificuldade para contratar e reter talentos. Havia entre os funcionários a suspeita de que candidatos externos aceitavam propostas da OpenAI apenas para usá-las como moeda de troca ao

negociar salários mais altos com o Google ou o DeepMind. A OpenAI precisava encontrar uma forma de se legitimar como organização de pesquisa.

Essas discussões eram recorrentes em almoços, em reuniões internas e em um documento estratégico da OpenAI intitulado "Brainstorming sobre o alcance da comunidade de pesquisa".[32] Na seção "Estratégia", o documento afirmava: "Trate explicitamente a comunidade de ML como uma parte interessada na comunicação", referindo-se à aprendizagem de máquina. "Mude nosso tom e nossas mensagens externas de forma que só os antagonizemos quando decidirmos intencionalmente fazê-lo." Já a seção "Política" reconhecia que uma reputação negativa na pesquisa científica poderia prejudicar seriamente a influência da OpenAI em Washington: "Para ter influência política em nível governamental, precisamos ser vistos como a fonte mais confiável em pesquisa de ML e AGI. O apoio e o respaldo generalizados da comunidade de pesquisa não são apenas necessários para ganhar essa reputação, mas também amplificam nossa mensagem."

A equipe de Clark elaborou um novo plano: um lançamento em etapas. Em vez de reter o GPT-2 permanentemente, a OpenAI passaria a divulgar progressivamente os modelos maiores que havia desenvolvido, em intervalos escalonados, e, caso tudo ocorresse bem, lançaria a versão completa com 1,5 bilhão de parâmetros. Essa abordagem, argumentou a equipe, permitiria que a OpenAI e outros pesquisadores observassem e lidassem gradualmente com quaisquer consequências que surgissem, além de dar tempo ao laboratório para firmar parcerias com outras organizações interessadas em estudar os riscos entre cada fase do processo.

Clark enfatizou para sua equipe a importância de construir um ecossistema por meio dessas parcerias. Trabalhar com instituições de prestígio ajudaria a promover uma cooperação mais ampla entre a indústria e a academia na abordagem dos riscos de segurança da IA. Além disso, essa estratégia garantiria maior adesão aos esforços da OpenAI para mudar as normas de divulgação de pesquisas e, ao mesmo tempo, melhoraria a reputação do laboratório. A equipe entrou em contato com pesquisadores de IA e segurança de algumas organizações selecionadas, concedendo-lhes acesso antecipado à versão completa do GPT-2 para testar seu potencial de uso em aplicações prejudiciais. Clark orientou a equipe a buscar "o maior apoio" possível de cada instituição, de modo que a OpenAI pudesse mencionar tanto os pesquisadores quanto suas organizações como parceiros.[33] Em seguida, a equipe preparou um livro branco apresentando a estratégia de lançamento e destacando essas parcerias. O documento buscava posicionar a OpenAI como líder no campo, listando exemplos de outras organizações que, após o GPT-2, também haviam se afastado da prática tradicional de divulgar imediatamente suas pesquisas.

O trabalho valeu a pena. Em pouco tempo, a iniciativa gerou discussões entre grupos do setor e *think tanks* políticos sobre a retenção de pesquisas como uma abordagem responsável para lidar com os riscos de segurança da

IA.[34] No fim de 2020, Clark e os irmãos Amodei deixariam a OpenAI para cofundar a Anthropic. Até lá, seu trabalho na OpenAI já teria consolidado sua influência e lançado as bases para suas ambições políticas em expansão.

A OpenAI começou a manter um roteiro para sistematizar sua pesquisa. Amodei tratava esse processo como um investidor, chamando-o de "um portfólio de apostas".[35] Ele e outros pesquisadores acompanhavam diferentes ideias dentro da área, cada uma nascida de distintas filosofias sobre como alcançar a inteligência artificial geral, e as desenvolviam por meio de experimentos em pequena escala. As ideias que se mostravam promissoras eram levadas adiante pela OpenAI; as que não apresentavam resultados satisfatórios eram abandonadas.

O projeto para vencer o campeonato do jogo de videogame *Dota 2* era uma área que Amodei acreditava não ter mais utilidade. A equipe havia derrotado os oponentes e atingido o objetivo em abril, o que ajudou a garantir o investimento da Microsoft e reforçou a confiança de alguns na estratégia de expansão da empresa. Com a missão concluída, o projeto foi encerrado e a equipe do *Dota 2,* dissolvida.

Onde Amodei viu um potencial contínuo foi no GPT-2.[36] O modelo representava uma aposta baseada no que o campo chamava de hipótese da "linguagem pura". Essa teoria sustenta que a linguagem é o principal meio de comunicação humana, o que implica que todo o conhecimento do mundo está, de alguma forma, registrado em texto. Assim, a IAG poderia emergir do treinamento de um algoritmo exclusivamente em grandes volumes de linguagem, sem a necessidade de outros tipos de dados. Essa ideia contrastava com a hipótese do "fundamento", segundo a qual o mundo físico, e nossa capacidade de percebê-lo e interagir com ele, é igualmente essencial para a inteligência. Nesse caso, a AGI só poderia surgir da combinação entre linguagem e percepção, como a visão computacional, e também da interação, seja por meio de um agente físico ou virtual executando ações no mundo real.

Nos documentos internos da empresa,[37] os pesquisadores analisavam os méritos das diferentes abordagens, enquanto a equipe de segurança de IA chegou a debater as virtudes da hipótese da "linguagem pura" usando analogias infelizes envolvendo pessoas com deficiência. As discussões expuseram como as tentativas de definir inteligência podem rapidamente se transformar em julgamentos preocupantes sobre quais grupos de pessoas seriam considerados mais ou menos inteligentes.

Sob o título "Alguns argumentos iniciais para a centralidade da linguagem", lia-se: "Alguma forma de linguagem é a diferença entre um ser humano selvagem e um ser humano na sociedade. Exemplo: Helen Keller."[38] Nas margens do documento, os pesquisadores de segurança de IA continuaram o debate, trocando comentários encadeados que apresentavam argumentos a favor e contra a hipótese da "linguagem pura".

"Além disso, as pessoas cegas são tão capazes quanto as pessoas com

visão", escreveu um pesquisador, argumentando que o "fundamento" por meio da visão parecia desnecessário.

Outro respondeu: "As pessoas cegas parecem estar em significativa desvantagem econômica", citando estatísticas da Federação Nacional dos Cegos que mostram que mais de 70% dos adultos com deficiência visual não trabalham em período integral.

Um terceiro interveio: "As pessoas cegas ainda são muito mais capazes do que os chimpanzés. Existem pessoas cegas bem impressionantes."

Muitos na OpenAI eram céticos quanto à hipótese da linguagem pura, mas o GPT-2 os fez reconsiderar. Treinar o modelo para prever a próxima palavra com precisão crescente impulsionou de forma significativa seu desempenho em outras tarefas de processamento de linguagem aparentemente pouco relacionadas. Tornou-se possível, até mesmo plausível, imaginar que um modelo GPT poderia desenvolver um conjunto mais amplo de capacidades seguindo esse caminho: aprimorando o treinamento e elevando ainda mais a precisão da previsão da próxima palavra.

Amodei começou a enxergar os modelos de linguagem em larga escala como — embora provavelmente não o único requisito para alcançar a AGI — talvez a rota mais rápida até ela. A situação se agravava porque a equipe de robótica enfrentava problemas constantes de hardware com sua mão robótica, produzindo a pior combinação possível: progresso lento e caro.

Contudo, havia um problema: se a OpenAI continuasse ampliando os modelos de linguagem, poderia intensificar os mesmos perigos sobre os quais havia alertado no caso do GPT-2. Amodei argumentou para a empresa — e Altman concordou — que isso não significava evitar essa direção. Na verdade, a conclusão era o oposto: a OpenAI deveria ampliar seu modelo de linguagem o mais rápido possível, afirmou Amodei, mas sem lançá-lo imediatamente. O GPT-2 havia mostrado como seria fácil para outros atores obter recursos de IA mais poderosos;[39] dois estudantes de pós-graduação, por exemplo, já haviam criado uma versão de código aberto do GPT-2 antes mesmo de a OpenAI liberar sua versão completa. Era apenas uma questão de tempo até que outras pessoas começassem a escalar ainda mais os modelos de linguagem.

Isso significava que, para garantir uma IAG benéfica, a melhor estratégia era a OpenAI dar um salto à frente e usar os prazos internos para descobrir como tornar seu modelo ampliado mais seguro. Quando chegasse o momento de revelar o modelo, esse refinamento adicional ajudaria a estabelecer novas normas de segurança em IA, da mesma forma que a retenção inicial do GPT-2 havia mudado os padrões de divulgação de pesquisa.

Com uma versão do GPT-2 disponível ao público, também havia indícios de que os perigos associados aos modelos de linguagem pura talvez não fossem tão graves. Pelo que a OpenAI sabia, ele não havia sido utilizado em campanhas coordenadas de desinformação em larga escala — e tais campanhas eram certamente problemas menores quando comparadas aos riscos existenciais potenciais da AGI.

"Obviamente, o uso indevido não é bom", afirmou Amodei, "porém um modelo de linguagem é muito menos poderoso do que uma AGI. Estou muito preocupado com a possibilidade de modelos de linguagem serem usados como arma para desinformação e coisas do tipo — isso realmente me assusta —, mas, ao mesmo tempo, é uma preocupação relativamente singular, clara e bem definida."

Na visão de dele, em outras palavras, ampliar o GPT-2 não era apenas o caminho potencialmente mais rápido para avançar rumo à IAG, mas também aquele cujos riscos ao longo do percurso permaneceriam relativamente limitados a aspectos que ele considerava gerenciáveis — como desinformação e informações equivocadas, em vez de catástrofes. Isso daria à OpenAI um campo de testes mais seguro para experimentar um sistema de IA poderoso, porém não tão poderoso, e permitiria resolver diversos problemas, incluindo o próprio processo de lançamento.

"O que é IAG? Como é a IAG?", perguntou Amodei. "Estamos na posição incômoda de não saber como ela é. Não sabemos quando isso vai acontecer. Então, buscamos coisas que não são IAG, mas que apresentam ao menos algumas das oportunidades e desafios da IAG. E a esperança é que, se conseguirmos lidar bem com essas coisas, então estaremos, de certa forma, prontos para os grandes campeonatos."

Era uma lógica que se apoiava em uma premissa específica: a de que a IAG, embora amorfa e incognoscível, também era inevitável. Durante anos, a OpenAI justificaria repetidamente seus comportamentos com variações desse mesmo argumento. Sob o espectro da chegada inevitável da IAG, a empresa considerava essencial continuar desenvolvendo modelos cada vez mais poderosos, tanto para se preparar quanto para preparar a sociedade. Mesmo que esses modelos carregassem os próprios riscos, a experiência que ofereciam para prevenir ou enfrentar um possível apocalipse da IA tornava esses riscos suportáveis.

À medida que o ChatGPT conquistava o mundo no início de 2023, um pesquisador chinês de IA compartilhou comigo uma análise lúcida que desmontava o argumento de inevitabilidade da OpenAI. O que a empresa fez jamais poderia ter acontecido em qualquer outro lugar além do Vale do Silício, disse ele. Na China, que rivaliza com os Estados Unidos em talento de IA, nenhuma equipe de pesquisadores e engenheiros, por mais impressionante que fosse, receberia 1 bilhão de dólares, muito menos dez vezes mais, para desenvolver uma tecnologia extremamente cara sem uma visão claramente articulada de como seria e para que serviria. Só depois do lançamento do ChatGPT foi que empresas e investidores chineses começaram a financiar com entusiasmo o desenvolvimento de modelos gigantescos, já com evidências suficientes de que poderiam recuperar o investimento por meio de aplicações comerciais.

Ao longo de minha apuração, cheguei a uma conclusão ainda mais surpreendente: nem no Vale do Silício outras empresas e outros investidores

passaram a direcionar somas ilimitadas para a escalabilidade antes do ChatGPT. Isso incluía o Google e o DeepMind, rivais originais da OpenAI. Foi especificamente a OpenAI, com suas origens bilionárias, sua inclinação ideológica singular e o impulso, a rede de contatos e o talento excepcional de Altman para arrecadar fundos, que criou a combinação perfeita para que sua visão particular surgisse e se impusesse. "Tenho a impressão de que Sam é a pessoa mais ambiciosa do planeta", diz um ex-funcionário. Em outras palavras, tudo o que a OpenAI fez foi o oposto de inevitável; os custos globais explosivos de seus modelos gigantes de aprendizagem profunda e a corrida arriscada que ela desencadeou na indústria para ampliar esses modelos até os limites do planeta só poderiam ter surgido exatamente no único lugar em que surgiram.

Para a demonstração a Gates, em abril de 2019, a OpenAI já havia ampliado o GPT-2 para algo um pouco maior. No entanto, Amodei não estava interessado em uma expansão modesta. Se o objetivo era aumentar a vantagem de tempo da OpenAI, o GPT-3 precisava ser o maior possível. A Microsoft estava prestes a entregar um novo supercomputador para a OpenAI como parte de seu investimento, com 10 mil GPUs Nvidia V100, na época, as mais poderosas do mundo para treinar modelos de deep learning. (O *V* homenageava o químico e físico italiano Alessandro Volta.) Amodei queria usar todos esses chips simultaneamente para criar o novo modelo de linguagem de grande escala.[40]

A ideia parecia, para muitos, nada menos que absurda. Até então, os modelos já eram considerados de grande escala quando treinados em algumas dezenas de chips. Nos melhores laboratórios acadêmicos do MIT e de Stanford, estudantes de doutorado consideravam um luxo ter dez chips. Em universidades fora dos Estados Unidos, como na Índia, os alunos tinham sorte quando conseguiam compartilhar um único chip com vários colegas, trabalhando com apenas uma fração de uma GPU para suas pesquisas.

Muitos pesquisadores da OpenAI duvidavam de que a ideia de Amodei chegaria a funcionar. Alguns também argumentavam que uma abordagem de ampliação mais gradual seria mais equilibrada, científica e previsível. Contudo, ele insistiu em sua proposta e contava com o apoio de outros executivos. Sutskever estava ansioso para testar sua hipótese de escalabilidade dos Transformers; Brockman queria continuar elevando o perfil da empresa; Altman pressionava para dar o maior salto possível. Pouco tempo depois, Amodei foi promovido a vice-presidente de pesquisa.

Nos bastidores, Altman também estava atento a outro fator: o investimento de 1 bilhão de dólares da Microsoft vinha acompanhado de expectativas de 1 bilhão de dólares; a OpenAI estava contra o relógio para entregar algo que justificasse aquele gasto. Enquanto Amodei via um modelo de linguagem maior como um pré-requisito necessário para a pesquisa de segurança em IA, Altman enxergava nele o potencial para cumprir a promessa

da OpenAI à Microsoft.

Nos meses seguintes, Amodei e Altman entrariam em conflito sobre como e quando lançar o GPT-3; Altman sairia vitorioso, pressionando para colocar o modelo no mundo em um cronograma acelerado. Anos antes do ChatGPT, essas duas decisões — ampliar drasticamente o tamanho do GPT-3 e lançá-lo depressa — mudariam o rumo do desenvolvimento da IA. Elas desencadeariam uma aceleração vertiginosa no avanço da IA, estimulando uma competição feroz entre empresas e países. Alimentariam uma expansão sem precedentes do capitalismo de vigilância e da exploração do trabalho. Pela quantidade colossal de recursos exigidos, consolidariam o desenvolvimento da tecnologia em um grau nunca visto, excluindo praticamente todo o restante do mundo de participar. Acelerariam o círculo vicioso em que universidades, incapazes de competir, perdem doutorandos e professores para a indústria, enfraquecendo a pesquisa acadêmica independente e marcando o início do fim da responsabilização. Ampliariam os impactos ambientais da IA a um nível que, na ausência de transparência ou regulamentação, nem especialistas externos, nem governos foram capazes de quantificar plenamente até hoje.

Entretanto, tudo isso ainda estava por acontecer. No outono de 2019, Amodei reuniu uma equipe, chamada Nest, composta principalmente de outros pesquisadores de segurança em IA, com a intenção de manter um controle rigoroso sobre o desenvolvimento do GPT-3 dentro da empresa. Com isso, a equipe iniciou sua ofensiva agressiva para escalar o modelo.

O GPT-3 era, na prática, o mesmo modelo que o GPT-2, alimentado com uma quantidade imensamente maior de dados e computação, tornando-se tão gigante que o resultado pareceria, para muitos, não apenas uma diferença de grau, mas uma diferença de natureza. No entanto, usar 10 mil chips trouxe novos problemas. Sempre havia uma pequena probabilidade de qualquer chip travar no meio do treinamento, da mesma forma que um laptop pode travar quando há muitas janelas abertas. Se um chip travasse, tudo travava, o que significava reiniciar o treinamento desde o começo. A probabilidade de uma única falha se multiplicava significativamente ao longo de 10 mil GPUs. Um erro desse tipo seria enormemente custoso — em dinheiro e em tempo —, já que a equipe da Nest esperava que o treinamento levasse, no mínimo, vários meses.

Para resolver o problema, a equipe precisava encontrar uma forma de garantir que o treinamento do modelo pudesse ser retomado exatamente de onde havia parado após qualquer interrupção. Também era necessário definir uma estratégia para distribuir o treinamento pelos 10 mil chips, um processo conhecido como *sharding*: seria melhor, por exemplo, dividir o modelo em dezenas, centenas ou milhares de partes, com cada uma delas treinando em clusters separados de GPUs antes de serem reunidas?

Havia ainda um desafio relacionado aos dados. Para obter o melhor desempenho, o tamanho do conjunto de dados precisava crescer

proporcionalmente ao número de parâmetros e à quantidade de computação. Se houvesse parâmetros de mais e dados de menos, o modelo poderia começar a regurgitar, palavra por palavra, os trechos presentes em seus dados de treinamento, tornando-se, na prática, inútil. No caso do GPT-2, Radford havia sido criterioso quanto ao que incluía no conjunto de dados. Ele coletou textos de artigos e sites compartilhados no Reddit que tivessem recebido ao menos três votos positivos na plataforma. Isso produziu um acervo de quarenta gigabytes com cerca de 8 milhões de documentos, ao qual ele deu o nome de WebText.[41]

Isso estava longe de ser suficiente para o GPT-3. Por isso, a equipe da Nest ampliou o conjunto de dados com uma varredura ainda mais extensa de links compartilhados no Reddit, além de uma coleta da Wikipédia em inglês e de um conjunto enigmático chamado Books2, cujos detalhes a OpenAI nunca divulgou.[42] Duas pessoas familiarizadas com o conjunto de dados me disseram que ele continha livros publicados extraídos do Library Genesis, um repositório obscuro on-line de livros e artigos acadêmicos pirateados. Em 2023, o Authors Guild, a principal organização de defesa dos direitos autorais nos Estados Unidos, e dezessete nomes, entre eles George R. R. Martin e Jodi Picoult, processariam a OpenAI e a Microsoft, alegando violação massiva de direitos autorais. A OpenAI responderia em março de 2024, afirmando que havia excluído esses conjuntos de dados e deixado de usá-los no treinamento depois do GPT-3.5, que àquela altura já estava obsoleto.[43]

Mesmo assim, ainda não eram dados suficientes. Então a equipe da Nest recorreu finalmente a um conjunto de dados público conhecido como Common Crawl, um enorme despejo de dados com petabytes, ou milhões de gigabytes, de textos, coletados regularmente de toda a internet.[44] Era uma fonte que Radford havia evitado de propósito, por ser de qualidade muito baixa.

Para tentar conter o lixo presente no conjunto, a equipe treinou um modelo de aprendizagem de máquina para identificar, dentro do Common Crawl, as amostras que mais se parecessem com artigos da Wikipédia. Caso se parecessem com a Wikipédia, a ideia era que teriam maior probabilidade de ter qualidade semelhante à da Wikipédia. A Nest também incluiu algumas amostras em outros idiomas além do inglês, embora elas representassem apenas 7% dos dados. Ainda assim, ao treinar o modelo, os pesquisadores atribuíram ao Common Crawl filtrado a menor prioridade. Em outras palavras, o GPT-2 havia representado o auge da qualidade dos dados; e, a partir dali, ela só caiu.

Quando chegou o momento de montar o conjunto de dados para o GPT-4, lançado dois anos depois, a pressão por quantidade corroeu ainda mais a qualidade. O filtro foi removido dos dados do Common Crawl, e a maior parte deles foi incorporada. Por meio de sua parceria com a Microsoft, a OpenAI também recebeu um download completo do GitHub, o repositório de código on-line pertencente à empresa. Quando isso ainda não foi

suficiente, funcionários da OpenAI passaram a coletar tudo o que conseguiam encontrar na internet: vasculharam links compartilhados no Twitter, transcreveram vídeos do YouTube e reuniram uma infinidade de outros conteúdos, incluindo blogs de nicho, despejos de dados já existentes e um site de armazenamento de texto chamado Pastebin. Tudo o que não tivesse um aviso explícito contra extração de dados era tratado como disponível para ser capturado.

Dentro do Google, alguns pesquisadores lamentavam a disposição da OpenAI em assumir riscos legais para coletar dados, algo que lhe conferia uma vantagem significativa. O Google era muito mais conservador em relação ao acesso e ao uso de dados e seguia um protocolo rigoroso para cumprir as regulamentações, incluindo a lei europeia de privacidade, conhecida como GDPR. Ironicamente, esse compromisso com a conformidade do Google acabava dando à OpenAI um acesso mais fácil aos dados da empresa do que o próprio Google tinha. Enquanto a OpenAI extraía e transcrevia rapidamente vídeos do YouTube, propriedade do Google, os pesquisadores do Google precisavam enfrentar uma burocracia interna extensa para cumprir o licenciamento restritivo do YouTube sobre o conteúdo enviado pelos usuários. A OpenAI não se preocupava — ou, no jargão das startups de tecnologia, "não se sentia sobrecarregada" — com essas exigências de conformidade. Era a velha mentalidade do Vale do Silício, em que fundadores e investidores defendem que startups podem e devem atuar nas áreas cinzentas da lei (como fizeram Airbnb, Uber e Coinbase) para provocar disrupção e revolucionar setores inteiros.

A decisão de reduzir as barreiras de qualidade — e, depois, eliminá-las por completo — teria efeitos amplos sobre a força de trabalho humana por trás dos sistemas de IA. Durante anos, a indústria de tecnologia contou com trabalhadores mal remunerados e em condições econômicas precárias para executar tarefas essenciais de preparação de dados para modelos de IA, como categorizar textos e rotular imagens. Logo após o GPT-3 normalizar o uso de conjuntos de dados gigantescos e de qualidade inferior, as demandas desse trabalho passaram de tarefas envolvendo conteúdo benigno para tarefas com conteúdo frequentemente perturbador, inclusive para fins de moderação, muito semelhante ao que já ocorria nas redes sociais. Essa moderação tornou-se necessária para evitar que os sistemas de IA generativa reproduzissem as partes mais vis de seus conjuntos de dados abrangentes — descrições ou representações de violência, abuso sexual ou automutilação — para centenas de milhões de usuários.

"Há uma grande mudança de paradigma na forma como você controla a produção desses modelos",[45] diz Ryan Kolln, CEO e diretor administrativo da Appen, uma plataforma que conecta empresas do Vale do Silício a profissionais que trabalham com dados. "No modelo tradicional de IA, você controla a saída restringindo as entradas" — o tipo de filtragem de dados que a equipe de Radford realizou —, "porque o sistema aprende apenas com os

exemplos que você fornece. O desafio da IA generativa é que as entradas são todo o *corpus* da humanidade. Então, você precisa controlar os resultados."

Em um artigo de 2023, Abeba Birhane e colaboradores apresentaram o conceito de "leis de escala do ódio"[46] para criticar a premissa de treinar modelos de aprendizagem profunda com dados não filtrados, aquilo que chamaram de "pântanos de dados". Eles analisaram dois conjuntos de dados públicos de imagens e textos usados para treinar geradores de imagens de código aberto, o LAION-400M e o LAION-2B-en, ambos extraídos do Common Crawl, com 400 milhões e 2 bilhões de imagens, respectivamente. Os autores mostraram que a quantidade de conteúdo odioso e abusivo aumentava conforme o conjunto de dados crescia, ampliando os comportamentos discriminatórios dos modelos treinados com base nele. Modelos treinados com os 2 bilhões de imagens, por exemplo, eram cinco vezes mais propensos a rotular rostos de homens negros como criminosos do que aqueles treinados com 400 milhões de imagens. Mais tarde naquele ano, um estudo de Stanford que examinou o LAION-5B, um conjunto com 5 bilhões de imagens usado para treinar o Stable Diffusion, descobriria que ele continha milhares de imagens de abuso sexual infantil comprovado ou suspeito.[47]

Entre as táticas usadas para controlar os resultados, a OpenAI contrataria trabalhadores no Quênia por, em média, menos de 2 dólares por hora para construir um filtro automatizado de moderação de conteúdo, uma revelação divulgada pela primeira vez por Billy Perrigo, correspondente da *Time*.[48] A empresa também empregaria mais de mil outros contratados em todo o mundo para realizar a aprendizagem por reforço a partir do feedback humano, o RLHF, técnica que havia desenvolvido para ensinar um agente de IA a dar saltos mortais, aplicada naquele momento a seus modelos de linguagem, o que incluía solicitar repetidamente respostas aos modelos e atribuir pontuações a essas respostas, numa tentativa de domar o comportamento do sistema o máximo possível.[49]

Hito Steyerl, artista e cineasta alemã que produziu um documentário sobre refugiados sírios envolvidos em trabalhos de rotulagem de dados, reforçou a crítica de Birhane nas observações que compartilhou comigo. Segundo ela, materiais psicologicamente prejudiciais se acumulam quando a vigilância em massa serve de base para a coleta de dados. Para enfrentar esse problema, é preciso voltar à raiz: questionar o que compõe esses dados e toda a premissa de sua coleta em larga escala e de forma indiscriminada.[50]

# II

CAPÍTULO 6

# ASCENSÃO

No início da carreira, Altman[1] percebeu que novos CEOs somente eram bem-sucedidos se "refundassem" a empresa. Na YC, ele fez isso com convicção ao assumir a presidência. Criou programas, incluindo um novo fundo, para expandir o apoio da incubadora para startups em diferentes estágios, e avançou para tecnologias complexas — aquelas que exigiam inovações científicas ousadas, como fusão nuclear, computação quântica e carros autônomos. Já uma marca de prestígio, a YC e sua esfera de influência cresceram de algumas centenas de empresas para milhares por ano, transformando-se em um centro gravitacional no Vale do Silício. "A coisa da qual mais me orgulho[2] é que realmente construímos um império", declarou Altman após deixar a presidência.

O fim de sua era na YC marcou o início de uma nova fase na OpenAI. Em março de 2019, ao se mudar em tempo integral para a OpenAI, ele rapidamente levou consigo a mesma mentalidade agressiva que havia aplicado à YC. Não queria que a empresa estivesse entre as principais organizações de IA do mundo; queria que fosse a única. Durante anos, por meio da incubadora e de outros fóruns, Altman ensinou outros fundadores a encararem o jogo das startups como uma competição na qual o vencedor leva tudo. Dizia que, se uma startup tivesse qualquer esperança de alcançar o sucesso, precisava agir de forma rápida e implacável para superar e continuar superando suas rivais.

O número mágico que ele costumava usar era o dez, derivado da estratégia de monopólio de Thiel. "Minha regra prática, um pouco louca[3] e um tanto arbitrária, é que se deve ter uma tecnologia que seja uma ordem de magnitude melhor do que a segunda melhor opção", dissera Thiel durante sua palestra em 2014 no curso de startups de Altman em Stanford. A Amazon, por exemplo, descobriu como vender dez vezes mais livros do que as livrarias físicas. O PayPal, sua própria empresa, descobriu como enviar pagamentos dez vezes mais rápido do que compensar cheques. "Você precisa ter algum tipo de melhoria que seja muito poderosa, talvez algo de uma ordem de magnitude, em alguma dimensão essencial", afirmara Thiel.

Altman usava o número redondo para tudo. Seu conselho era: as startups não apenas precisavam entrar no mercado com uma tecnologia dez vezes melhor, mas também melhorá-la dez vezes a cada nova geração. A velocidade com que atingiam cada nova geração era outra variável crucial que poderia defini-las ou destruí-las. "Se o seu ciclo de renovação[4] é de uma semana e o do seu concorrente é de três meses, você irá fazê-lo engolir poeira", disse em 2017 a uma turma de aspirantes a empreendedores.

Na YC, Altman pressionou seus sócios a continuarem aumentando em dez vezes o número de empresas financiadas. "E, com o tempo, descobriremos[5] como conseguir mais dez vezes, e depois mais dez vezes além disso", disse posteriormente sobre sua estratégia em um evento. "Um dia, financiaremos todas as empresas do mundo." "Sam foi a primeira pessoa[6] que ouvi dizer que, por conta do trabalho que os fundadores originais haviam feito e por causa da marca que a YC havia criado, éramos de fato um monopólio real nesse espaço", disse Geoff Ralston, seu sucessor na incubadora.

Na OpenAI, Altman planejava usar a mesma estratégia. Em um memorando enviado[7] à empresa no final de 2019 a fim de articular sua visão de longo prazo, enfatizou que a OpenAI precisava "ser a número um" em quatro categorias até o final do ano seguinte: resultados técnicos, capacidade de computação, dinheiro (para adquirir mais capacidade de computação) e preparação, o que significava a segurança e a proteção da organização, assim como sua resiliência em situações de alto estresse.

Segundo ele, o mais importante desses pontos era o primeiro. Caso a OpenAI quisesse ter uma chance de cumprir sua missão, precisava desenvolver primeiro uma IAG benéfica, ou ser uma líder tão dominante que ainda fosse capaz de moldar o desenvolvimento da IAG. "Embora, em teoria, pudéssemos desacelerar o trabalho de aprimoramento de capacidades", escreveu, referindo-se ao avanço dos resultados técnicos. "Dado o ritmo de progresso que outras pessoas estão alcançando, provavelmente seremos obrigados a avançar muito depressa no desenvolvimento técnico se quisermos ter grande influência sobre a IAG." Isso apenas se tornaria cada vez mais verdadeiro à medida que mais e mais concorrentes percebessem a estratégia da OpenAI e entrassem nesse espaço.

"Ainda precisamos de muitos outros saltos dez vezes maiores para chegarmos à IAG", acrescentou mais tarde ao memorando. "Devemos sempre trabalhar em busca de resultados dramáticos, não de melhorias incrementais."

Várias outras considerações eram cruciais para essa fórmula de sucesso. Seria fundamental para a OpenAI manter a satisfação da Microsoft, a fim de preservar a liderança em capacidade de computação. Caso a OpenAI fosse bem-sucedida, a gigante tecnológica havia concordado em oferecer muito mais do que 1 bilhão de dólares. "Queremos que a Microsoft seja nossa principal parceira durante todo o processo", disse Altman. "Eles têm a capacidade de nos fornecer, ao menos pelos próximos cinco anos, os supercomputadores mais poderosos do mundo." Isso significava se afastar dos tempos em que a OpenAI

cultivava um ambiente acadêmico de pesquisa livre e passar a focar esforços de comercialização direcionados para gerar benefícios à Microsoft. Caso a empresa quisesse realizar outros projetos de pesquisa, então teria os recursos necessários. "Parafraseando aquela famosa citação da Disney", escreveu Altman, "devemos ganhar mais dinheiro para podermos realizar mais pesquisas, não realizar mais pesquisas para podermos ganhar mais dinheiro."

Além disso, a empresa precisava começar a reduzir a transparência. "O risco de *infohazard* — ou risco de informação — ao falar de IAG continuará aumentando à medida que avançamos mais", argumentou Altman. Era hora de restringir as publicações de pesquisas e as implantações de modelos, adotar uma política de confidencialidade mais rígida e revelar o progresso apenas em habilidades específicas, em vez de avanços mais amplos na IA geral. Separadamente, todos também precisavam começar a agir sob a suposição de que "toda decisão que tomamos e toda conversa que temos acabam sendo investigadas e publicadas na primeira página do *New York Times*".

Dito isso, "ainda parece muito importante que o mundo pense que estamos vencendo em alguma coisa", disse. Isso faria com que "pessoas que são mundialmente influentes" estivessem mais "dispostas a se esforçar bastante para nos ajudar" e faria com que formuladores de políticas globais "no nível de presidentes ou de seus representantes" recorressem à OpenAI "em busca de respostas quando precisarem tomar grandes decisões". Nesse sentido, "provavelmente deveríamos planejar lançar ao menos uma demonstração de progresso que exceda as expectativas a cada ano".

Por fim, a empresa precisava começar a agir com mais seriedade e unidade. Altman incluiu uma citação de Hyman G. Rickover, um almirante da Marinha dos EUA conhecido como o "pai da marinha nuclear" por seu trabalho na construção dos primeiros submarinos movidos a energia nuclear do mundo. Era uma citação que Altman havia mandado pintar nas paredes do escritório nos primeiros dias da OpenAI:

> Acredito que é dever de cada um de nós agir como se o destino do mundo dependesse de nós. É verdade que uma única pessoa, sozinha, não pode fazer todo o trabalho. Cada um de nós é obrigado a colocar suas capacidades individuais e independentes a serviço de uma ampla gama de preocupações humanas. É com essa convicção que enfrentamos diretamente nosso dever para com a prosperidade. Devemos viver pelo futuro da raça humana, não pelo nosso próprio conforto ou sucesso.

"Construir uma IAG que beneficie a humanidade é talvez o projeto mais importante do mundo", escreveu Altman abaixo da citação no documento. "Devemos colocar a missão acima de quaisquer preferências pessoais. Questões de baixa importância devem ser de baixo drama, para que possamos reservar nossa capacidade de alto drama para as questões de alta importância

(e haverá muitas delas)."

O drama, na verdade, já estava pronto para acontecer. Várias pequenas divergências que haviam surgido pela empresa começavam a se aglutinar em conflitos maiores. Se antes começaram rapidamente a se chamar de amigos, Brockman e os irmãos Amodei passaram a discordar em uma lista crescente de questões. Entre elas, a decisão de Dario de dar menos prioridade ao trabalho com o *Dota 2*, que havia frustrado Brockman, o qual acreditava que ele não levava suas contribuições a sério. Enquanto o *Dota 2* já tinha sido o projeto que mais demandava capacidade de computação, Brockman também se irritava com a centralização de tal capacidade por parte de Amodei para o trabalho do Nest com o GPT-3. Os irmãos Amodei, por sua vez,[8] consideravam Brockman uma pessoa difícil de lidar e não estavam dispostos a deixá-lo participar do desenvolvimento de seus modelos de linguagem.

As tensões criaram uma divisão entre as lideranças que, aos poucos, se entendeu às pessoas leais de cada lado dentro da empresa. Durante o projeto *Dota 2*, Brockman havia criado um vínculo quase familiar com alguns membros de sua equipe devido às longas horas de trabalho, ao alto estresse e a uma viagem improvisada ao Havaí, aproximando-se principalmente de Jakub Pachocki e Szymon Sidor, dois cientistas poloneses que eram colegas de apartamento e melhores amigos. As equipes de segurança de IA de Amodei — sobretudo os membros principais da equipe do Nest — formavam outro grupo, unidas por uma preocupação em comum, mas em diferentes graus, com a possibilidade de uma IA descontrolada e com riscos existenciais ou outros riscos extremos. Mantinham seu trabalho isolado do restante da empresa, criando canais privados no Slack e documentos inacessíveis até mesmo a outros executivos. Isso frustrou muitas pessoas além de Brockman, que também se sentiam deixadas de lado por conta da redução dos recursos computacionais disponíveis e da perda de visibilidade sobre as principais pesquisas da empresa.

Enquanto isso, o grupo de segurança de IA de Amodei também começava a se incomodar com alguns comportamentos de Altman. Pouco depois de o acordo com a Microsoft ser firmado, muitos deles ficaram chocados ao descobrir a extensão das promessas que Altman fizera à contraparte acerca das tecnologias às quais a empresa teria acesso em troca de seu investimento. Os termos do acordo não correspondiam ao que haviam entendido a partir das explicações do CEO. Eles temiam que, caso surgissem problemas de segurança de IA nos modelos da OpenAI, tais compromissos dificultariam muito, se não impossibilitariam, que sua instalação fosse impedida. Sendo assim, o grupo de Amodei começou a ter sérias dúvidas sobre a honestidade de Altman.

"Somos todos pessoas pragmáticas", disse um integrante do grupo. "Obviamente estamos levantando dinheiro; criaremos produtos comerciais. Se você for alguém que faz inúmeros acordos como Sam, pode parecer muito

razoável pensar: 'Certo, vamos firmar um acordo, trocar uma coisa, e depois trocamos a próxima.' No entanto, se você for como eu, pensa: 'Estamos trocando algo que não entendemos totalmente.' Parece que isso nos deixa em uma posição desconfortável."

Tudo isso acontecia em meio a uma crescente paranoia em relação a diversas questões dentro da empresa. No grupo de segurança de IA, ela girava em torno do que consideravam evidências cada vez mais fortes de que sistemas de inteligência artificial poderosos e desalinhados poderiam levar a resultados catastróficos. Uma experiência particularmente bizarra havia deixado vários deles um tanto apreensivos. Em 2019, com um modelo treinado após o GPT-2 e com aproximadamente o dobro de parâmetros, um grupo de pesquisadores começou a desenvolver o trabalho de segurança de IA que Amodei desejava: testar o aprendizagem por reforço a partir de feedback humano como uma forma de orientar o modelo a gerar conteúdo alegre e positivo, afastando-o de qualquer coisa ofensiva.

Entretanto, certa noite, um pesquisador fez uma atualização que continha um único erro de digitação em seu código antes de deixar o processo de RLHF rodando da noite para o dia. Esse erro era importante: um sinal de menos trocado por um sinal de mais, o que fez o processo funcionar ao contrário, levando o GPT-2 a gerar *mais* conteúdo ofensivo em vez de menos. Na manhã seguinte, o erro já havia causado estragos, e o GPT-2 completava todos os comandos com linguagem extremamente obscena e sexualmente explícita. Foi hilário — e preocupante também. Após identificar o erro, o pesquisador enviou uma correção para a base de código da OpenAI com um comentário: "Que tal não criarmos um minimizador do serviço?"

Parcialmente motivados pela percepção de que apenas o aumento da escala já poderia gerar novos avanços em IA, muitos funcionários também se preocupavam com o que aconteceria se outras empresas descobrissem o segredo da OpenAI. "O segredo de como nossas coisas funcionam cabe escrito em um grão de arroz", diziam uns aos outros, referindo-se a uma única palavra: *escala*. Pelo mesmo motivo, havia o receio de que capacidades poderosas parassem em mãos de agentes mal-intencionados. A liderança alimentava esse medo, mencionando frequentemente a ameaça da China, da Rússia e da Coreia do Norte, e enfatizando a necessidade de que o desenvolvimento da IAG permanecesse sob o controle de uma organização dos EUA. Às vezes, isso irritava os funcionários que não eram norte-americanos. Durantes os almoços, questionavam: "Por que tinha que ser uma organização dos EUA?", lembrou um ex-funcionário. "Por que *não* uma da Europa? Ou da China?"

Durante essas intensas discussões filosóficas sobre as implicações de longo prazo da pesquisa em IA, muitos funcionários constantemente retornavam às analogias iniciais de Altman sobre a OpenAI e o Projeto Manhattan. A empresa estaria de fato construindo o equivalente a uma arma nuclear? Era um contraste estranho com a cultura corajosa e idealista que a organização

havia criado até então, em grande parte acadêmica. Às sextas-feiras, eles relaxavam após uma longa semana em noites regadas a música e vinho, desfrutando do som suave de um grupo de colegas tocando o piano do escritório até altas horas da noite.

A mudança de foco e gravidade fez com que alguns se sentissem inquietos, ficando mais ansiosos em relação a incidentes aleatórios e aparentemente não relacionados. Em certa ocasião, um jornalista seguiu alguém dentro do estacionamento fechado para conseguir acesso ao prédio; em outra, um funcionário encontrou um pen drive não registrado, causando preocupação sobre a possibilidade de conter arquivos maliciosos, um vetor comum de ataque, e de ser uma tentativa de violação de segurança cibernética. Após examiná-lo em um computador sem conexão com a internet, viram que o pen drive não continha nada. Ao menos duas vezes, Amodei também usou um computador sem conexão para redigir documentos estratégicos críticos, conectando a máquina diretamente a uma impressora para circular apenas cópias físicas. Ele era paranoico quanto à possibilidade de agentes estatais roubarem os segredos da OpenAI e construírem os próprios modelos de inteligência artificial poderosos para fins maliciosos.

"Ninguém estava preparado para ter essa responsabilidade", lembrou um funcionário. "As pessoas chegavam a perder o sono."

O próprio Altman era paranoico[9] quanto ao vazamento de informações. Preocupava-se particularmente com a equipe da Neuralink, com quem a OpenAI continuava a compartilhar um escritório, na época com mais apreensão após a saída de Musk. Ele também se preocupava com Musk, que tinha um aparato de segurança extenso, incluindo motoristas particulares e seguranças. Altman, em determinado momento, ciente da diferença de recursos, encomendou secretamente uma auditoria eletrônica de contravigilância na tentativa de vasculhar o escritório em busca de possíveis dispositivos de escuta que Musk pudesse ter deixado para espionar a OpenAI.

Para os funcionários, Altman falava da hipótese de adversários dos EUA avançarem na pesquisa de inteligência artificial mais rapidamente do que a OpenAI, para justificar por que a empresa precisava ser cada vez menos aberta, enquanto trabalhava o mais rápido possível. "Devemos nos responsabilizar por um bom resultado para o mundo", escreveu em seu documento sobre a visão da empresa. "Por outro lado, se um governo autoritário desenvolver uma IAG antes de nós e usá-la indevidamente, também teremos falhado em nossa missão — quase certamente precisamos apressar nossos avanços técnicos para que ela seja bem-sucedida."

Altman começou a pressionar a equipe em relação à segurança. Os executivos debatiam sobre onde traçar o novo limite: a OpenAI deveria agir mais como uma empresa da Fortune 500, protegendo tecnologias proprietárias, ou mais como uma operação governamental, protegendo segredos de Estado extremamente confidenciais? No mínimo, eles concordavam que havia a

necessidade de proteger os pesos dos modelos — a informação principal que poderia ser utilizada para replicar as versões integralmente treinadas dos modelos de aprendizagem profunda da OpenAI. Seria gravíssimo se fossem roubados, pois poderiam tanto fortalecer agentes mal-intencionados quanto prejudicar a vantagem competitiva da empresa.

A princípio, sem equipe de segurança formal, Altman designou um membro do time de infraestrutura, que era responsável por tudo, desde as GPUs até a internet do escritório da empresa, para pensar em soluções que evitassem o roubo dos modelos — não apenas por espiões corporativos ou patrocinados pelo estado, mas também por funcionários da própria OpenAI. Em cibersegurança, proteger-se contra a "ameaça interna" é relativamente comum. Funcionários internos, caso tivessem a intenção, poderiam sabotar ou roubar a propriedade intelectual da OpenAI, e até mesmo ser ludibriados para entregá-la. Em privado, Altman reconheceu, após a questão ser levantada, que alguém como Sutskever poderia ser vulnerável à segunda possibilidade. O cientista-chefe era um alvo óbvio para agentes mal-intencionados: o arquétipo do cientista genial que não era muito esperto no dia a dia, mas ocupava posição elevada na organização e tinha acesso privilegiado às informações.

Sutskever também tinha as próprias paranoias. Como cientista de destaque no mundo intelectual e socialmente inepto da pesquisa de IA, já havia lidado com fãs obsessivos e atos de perseguição. Mais de uma vez, estranhos tentaram entrar sorrateiramente no escritório da OpenAI apenas para vê-lo. Assim como Amodei, ele também se preocupava com o poder da inteligência artificial de atrair a atenção de governos inescrupulosos e se perguntava se aqueles ansiosos por seus conselhos seriam secretamente agentes estrangeiros. Comentava com colegas o que deveria fazer caso sua mão fosse decepada para ser usada em um scanner de biometria palmar e desbloquear os segredos da empresa. Queria contratar menos pessoas e manter uma equipe pequena, a fim de reduzir os riscos de infiltração. Com Jakub Pachocki e Szymon Sidor, propôs construir uma instalação de contenção segura, um bunker com um computador isolado da rede, que abrigaria todos os pesos dos modelos da OpenAI e impediria que outros os roubassem. A ideia, que não fazia sentido na prática, pois os modelos precisavam ser treinados primeiro nos servidores da Microsoft, nunca saiu do papel.

Longe dos olhares da maioria dos funcionários, a segurança digital aumentou com a instalação de softwares de monitoramento corporativo, e, em paralelo, a segurança física também foi intensificada. Os portões do estacionamento do escritório foram reforçados; dentro das instalações, diversas portas com fechadura eletrônica foram programadas com "senhas emergenciais", códigos especiais que podiam ser digitados para acionar um alarme secreto, alertando a equipe de segurança responsável sobre uma ameaça presencial. Foram solicitados orçamentos a fornecedores a fim de reforçar uma sala de servidores para que resistisse a tiros de metralhadora, embora a ideia tenha sido posteriormente abandonada.

No memorando sobre a visão da empresa, Altman observou as divergências que estavam tomando conta da OpenAI devido ao aumento de estresse. "Temos (ao menos) três clãs entre nós — para caricaturá-los, digamos, que há o de Pesquisa Exploratória, o de Segurança e o de Startup." O primeiro estava concentrado em avançar as capacidades da inteligência artificial; o segundo, em focar a responsabilidade; e o terceiro, em agir rapidamente e fazer as coisas acontecerem.

Segundo Altman, cada um desses clãs tinha valores significativos que a empresa precisava preservar: o "vamos perseguir novas ideias importantes mesmo que falhemos diversas vezes" do clã de Pesquisa Exploratória; o "teremos um compromisso inabalável de fazermos a coisa certa" do clã de Segurança; e o "vamos encontrar um jeito de fazer acontecer" do clã de Startup. "Precisamos continuar evitando guerras tribais", disse ele. "Para termos sucesso, necessitamos que esses três clãs se unam como uma única tribo — mantendo as forças individuais de cada um — para trabalhar na direção de uma IAG que beneficie a humanidade ao máximo."

Embora Altman nunca tenha mencionado nomes, os funcionários perceberam o que estava subentendido. Sutskever era a face do clã de Pesquisa Exploratória; Amodei e seu grupo de segurança de IA, focado em riscos extremos, constituíam o clã de Segurança; e Brockman era o campeão do clã de Startup. Pouco depois, a pandemia aconteceu, e todos começaram a trabalhar remotamente, tornando muito mais fácil para os grupos se isolarem uns dos outros.

Amodei pressionou sua equipe para avançar depressa. Como haviam feito[10] com o GPT-2, eles treinaram modelos progressivamente maiores até chegarem a um que fosse composto de 10 mil GPUs com 175 bilhões de parâmetros, nomeando-os em ordem alfabética em homenagem a cientistas: *ada*, para o menor modelo, em referência à matemática inglesa Ada Lovelace, amplamente reconhecida como a primeira programadora de computadores; *babbage*, para honrar o inventor inglês Charles Babbage, que concebeu o primeiro computador digital para o qual Lovelace propôs seu programa; *curie*, agraciando a física e química polonesa-francesa Marie Curie, a primeira mulher a receber o Prêmio Nobel e ainda a repetir o feito; e *davinci*, aludindo a Leonardo Da Vinci. O objetivo do exercício era tanto continuar validando se as leis de escalabilidade ainda se mantinham em uma escala fundamentalmente maior quanto, de maneira mais prática, trabalhar gradualmente para superar os desafios de hardware e dados em cada novo nível. Com frequência, a equipe do Nest apresentava à empresa atualizações sobre seu progresso, gerando crescente empolgação. "É difícil expressar o quão insano foi ver aquilo", lembrou um pesquisador. "Nunca tinha visto nada parecido na minha vida."

Em paralelo, Altman e Brockman desenvolveram um plano de comercialização. No fim de janeiro de 2020, Brockman começou a escrever as

primeiras linhas de código para uma interface de programação de aplicativos, ou API, do GPT-3. Ela daria a empresas e desenvolvedores acesso às capacidades do modelo sem lhes conceder abertura aos respectivos pesos, permitindo que incorporassem a tecnologia nos próprios produtos voltados ao consumidor. A OpenAI, então, se dividiu em duas seções. Mira Murati foi promovida a vice-presidente de uma nova subdivisão da área de Produtos, responsável pela supervisão da API e estratégia de comercialização. Sob sua liderança, Peter Welinder, que vinha chefiando a equipe de robótica, passou a comandar a área de produtos; Fraser Kelton, cofundador de uma startup de IA adquirida pela Airbnb, e Katie Mayer, que havia trabalhado na Leap Motion, foram contratados para, respectivamente, gerenciar as novas equipes de produto e engenharia. Por associação, todos que não faziam parte da área de Produtos passaram a compor a subdivisão de Pesquisa.

A divisão aprofundou a fratura que Altman havia identificado. A formação da subdivisão da área de Produtos trouxe um grupo pequeno, mas crescente, de pessoas vindas de outras startups, fortalecendo o respectivo clã. Enquanto o de Pesquisa Exploratória encarava essa mudança com certa hesitação, indeciso se a OpenAI se tornaria apenas mais uma empresa de produtos do Vale do Silício, ela despertou uma oposição cada vez mais intensa de Amodei e de seu clã de Segurança, também situado na mesma subdivisão.

Para muitos dentro da equipe de Segurança, o lançamento do GPT-3 em tão pouco tempo através de uma API, ou de qualquer outro meio, minava justamente a vantagem de tempo — o propósito central da aceleração das escalas — que a OpenAI teria para aperfeiçoar a defesa do modelo. No entanto, a subdivisão da área de Produtos, cuja finalidade era encontrar soluções iniciais para lucrar com as tecnologias da OpenAI, o que em sua visão exigia lançá-las em curto prazo, discordava. A API, como eles viam, também garantia à empresa mais controle do mecanismo de qualquer estratégia de lançamento, permitindo que a OpenAI fosse seletiva em relação a quem conceder acesso e coletasse dados valiosos para entender como o modelo poderia ser usado ou abusado pelos usuários. Nas reuniões gerais, Altman conciliava ambos os lados: a API acabaria ajudando cada grupo a alcançar o que desejava; gerar receita permitiria que a empresa investisse ainda mais em pesquisa de segurança de IA.

Quando o treinamento do GPT-3 foi concluído, os funcionários começaram a experimentar o modelo internamente, testando os limites de suas capacidades e fazendo ajustes na primeira versão da API. A empresa realizou um hackathon, um evento no qual os funcionários elaboraram ideias sobre diferentes aplicações. Contudo, a cada novo protótipo, as tensões pioravam. Enquanto a subdivisão da área de Produtos e muitos integrantes do clã de Pesquisa Exploratória assistiam às demonstrações com entusiasmo crescente, vários funcionários da equipe de Segurança as viam como mais uma prova de que lançar o modelo sem testes abrangentes e pesquisas adicionais poderia trazer consequências devastadoras.

Uma capacidade se mostrou particularmente polarizadora: as habilidades de geração de código do GPT-3. Isso não fazia parte das intenções da equipe do Nest, mas, ao coletar links no Reddit e utilizar o Common Crawl como base de dados de treinamento, eles haviam capturado trechos aleatórios de código publicados por engenheiros em diversos fóruns on-line para tirar dúvidas ou compartilhar dicas, fazendo o modelo desenvolver uma facilidade maior com linguagens de programação. Muitos da equipe de Pesquisa Exploratória ficaram animados com isso, assim como da subdivisão da área de Produtos. Além de representar um marco técnico impressionante,[11] também tinha potencial como ferramenta para acelerar a produtividade da empresa na pesquisa de IA e para tornar o GPT-3 um produto mais atraente. Pelo mesmo motivo, alguns membros da equipe de Segurança entraram em pânico. Se um sistema de inteligência artificial pudesse usar as próprias habilidades para gerar códigos e se modificar, isso poderia acelerar a chegada de capacidades mais poderosas, aumentar o risco de subversão do controle humano e ampliar as chances de riscos extremamente prejudiciais ou existenciais.

Sutskever e Wojciech Zaremba, um dos membros fundadores que Musk havia pressionado durante uma reunião, formariam mais tarde uma equipe para a criação de um modelo projetado especificamente para gerar códigos. No entanto, durante uma reunião para iniciar o projeto, ambos descobriram que Amodei já tinha os próprios planos para desenvolver um modelo de geração de código e não via necessidade de unir esforços. Apesar de suas preocupações, ele acreditava, como no caso do GPT-3, que a melhor forma de mitigar os possíveis danos da geração de código era simplesmente construir o modelo mais rápido do que qualquer outra pessoa — até mesmo que as demais equipes da OpenAI, as quais, de acordo com sua visão, não priorizariam a segurança da IA — e se valer dessa vantagem de tempo para realizar pesquisas sobre a redução de riscos do modelo. Para a confusão de outros funcionários, ambas as equipes continuaram unindo esforços para o mesmo objetivo. "Para quem via de fora, aquilo tudo parecia uma versão maluca de *Game of Thrones*", disse um pesquisador.

O impasse em torno do lançamento do GPT-3 por meio da API continuou até o final da primavera. O clã de Segurança continuou a pressionar para que se tivesse extrema cautela devido aos receios de acelerar riscos extremos da IA, argumentando que a empresa deveria adiar o lançamento o máximo possível. A subdivisão da área de Produtos prosseguia com os preparativos para o lançamento da API, defendendo que a melhor maneira de aprimorar o modelo era permitir que ele tivesse contato com o mundo real. Ao mesmo tempo, surgiram novas preocupações de um terceiro grupo de funcionários, apreensivos com o impacto que habilidades extraordinárias de geração de texto poderiam causar em meio a grandes turbulências políticas, sociais e econômicas nos EUA. Em maio de 2020, a pandemia já havia criado um aumento na taxa de desemprego[12] mais rapidamente do que durante a Grande Recessão. No mesmo mês, Derek Chauvin, um policial de

Minneapolis, assassinou George Floyd, um homem negro de 46 anos, desencadeando enormes protestos do movimento Black Lives Matter pelo país e pelo mundo afora. A equipe também estava preocupada com a eleição presidencial que se aproximava.

Entretanto, dentro da OpenAI, começaram a circular rumores de que o Google poderia em breve lançar o próprio grande modelo de linguagem. A possibilidade era plausível. O Google já havia publicado[13] uma pesquisa no início do ano sobre um novo chatbot chamado Meena, que tinha como base um grande modelo de linguagem com 1,7 vez mais parâmetros do que o GPT-2. A empresa poderia facilmente estar trabalhando para escalar o modelo para algo próximo ao tamanho do GPT-3. Os rumores selaram a decisão de lançar a API: se um modelo igualmente grande logo existiria no mundo, o clã de Segurança não via muitos motivos para reter o GPT-3.

Em junho, a empresa anunciou a API e criou um formulário para que as pessoas solicitassem acesso antecipado, priorizando grandes empresas que, de acordo com a subdivisão da área de Produtos, poderiam ser confiáveis para lidar com a tecnologia com responsabilidade. Além disso, manteve uma extensa planilha para que os funcionários adicionassem o nome de qualquer pessoa que quisesse "furar a fila", incluindo familiares, amigos e suas celebridades preferidas.

O suposto modelo do Google nunca se tornou realidade. A gigante tecnológica de fato havia começado a trabalhar em uma versão maior do que a Meena, conhecida como LaMDA, para produzir um chatbot melhor — porém ainda era modestamente menor do que o GPT-3, e a empresa acabaria optando por não lançá-lo até depois do lançamento do ChatGPT. Os executivos do Google determinaram[14] que o LaMDA não atendia aos padrões éticos de IA da empresa. Alguns funcionários também temiam um repeteco do famoso escândalo da Microsoft: em 2016, ela lançou um chatbot de inteligência artificial chamado Tay, que rapidamente se tornou racista e misógino e passou a expressar apoio a Hitler após usuários induzirem o bot a repetir coisas impróprias e ofensivas. O lançamento da API do GPT-3 não seria a última decisão que a OpenAI tomaria para divulgar sua tecnologia com base no medo exacerbado da concorrência.

Assim como o ChatGPT tornaria a OpenAI instantaneamente conhecida pelo público, o GPT-3 representou esse momento dentro dos círculos de IA e tecnologia. No fim de 2022, o ChatGPT adicionaria melhorias e recursos essenciais à experiência do GPT-3 que o transformariam em um produto viral global, incluindo uma interface amigável ao consumidor, capacidades conversacionais, mais mecanismos de segurança e uma versão gratuita. No entanto, muitas das capacidades centrais que o público em geral experimentaria com o chatbot na época, os desenvolvedores já estavam vivenciando com a API dois anos antes, em 2020. Com a mesma admiração e surpresa, eles não conseguiam acreditar no que viam.

As capacidades do GPT-3 estavam muito além de tudo o que o GPT-2 havia exibido. Nunca alguém na pesquisa ou na indústria tinha visto uma tecnologia capaz de gerar ensaios, roteiros e códigos aparentemente com a mesma destreza. Esse tipo de flexibilidade para executar diversas tarefas já era, por si só, extremamente impressionante do ponto de vista técnico — modelos de linguagem anteriores, no geral, tinham apenas aptidão para realizar a única tarefa para a qual haviam sido treinados. No entanto, muitos acreditavam que o GPT-3, ainda mais extraordinário, estava começando a exibir outra característica, uma que havia sido cobiçada por muito tempo na área: a rápida generalização. Mostrar ao modelo alguns exemplos de uma nova tarefa que se desejava que ele executasse era o bastante para colocá-lo em funcionamento.

Na NeurIPS daquele ano,[15] o artigo da OpenAI explicando seu trabalho sobre o modelo venceu um dos principais prêmios de pesquisa, surpreendendo os funcionários e consolidando o status do laboratório como uma organização pioneira. O efeito foi exatamente como a equipe de liderança havia previsto. O prestígio adquirido pela OpenAI facilitou o recrutamento e a retenção de talentos, impulsionado significativamente pelo capital que a OpenAI LP levantou, permitindo que a empresa enfim competisse com o Google e a DeepMind no que se referia a salários.

Em outubro de 2020, com o reconhecimento crescente da OpenAI, Altman contratou Steve Dowling, um executivo experiente que havia liderado a área de comunicações da Apple, para ser o novo vice-presidente de comunicações da empresa. Também colocou Dowling no comando das relações governamentais, enfatizando a importância de educar formuladores de políticas sobre a IA e torná-los conscientes das capacidades que estavam por vir. Após a saída de Jack Clark, Dowling trouxe Anna Makanju, uma ex-conselheira altamente respeitada durante a administração Obama, que também havia trabalhado com políticas no Facebook e no Starlink de Musk, para assumir a área de políticas e assuntos globais.

Ansiosa para aproveitar o impulsionamento do GPT-3, a subdivisão da área de Produtos elaborou ideias para desenvolver e ampliar sua estratégia de comercialização. No entanto, aparentemente a cada passo, o clã de Segurança continuava a oferecer resistência. Para eles, que ainda lidavam com o lançamento precipitado do GPT-3, a melhor forma de remediar a liberação prematura não era reproduzi-lo ainda mais, mas primeiro resolver as deficiências do modelo o mais rápido possível. A versão em vigência na API não contava com nenhum tipo de filtragem de moderação de conteúdo, assim como seus resultados não haviam sido refinados por meio de aprendizagem por reforço a partir de feedback humano. Em reuniões, ambos os campos procuravam encontrar um meio-termo, mas, em vez disso, falavam em círculos intermináveis. Em certo momento, Welinder,[16] que se tornaria vice-presidente de produtos, comentou amargamente que toda conversa parecia a reencenação de um manual de inteligência dos EUA de 1944 sobre sabotagem não violenta. Uma seção do folheto, a qual se tornou pública em 2008, lista

instruções simples sobre como desestabilizar e minar a produtividade de uma organização, incluindo:

- Fale o mais constantemente possível e por um longo período.
- Levante questões irrelevantes com a maior frequência possível.
- Discuta sobre a exatidão das palavras usadas em comunicações, atas e resoluções.
- Remeta-se a assuntos decididos na reunião anterior e tente reabrir a questão da conveniência de tal decisão.
- Faça o máximo de perguntas que puder.

A animosidade também permeava o ambiente fora das reuniões. Para a equipe da subdivisão da área de Produtos, parecia que todos os canais de comunicação digital estavam sendo transformados em campos de batalha. Um post de alguém da área de produtos no Slack podia desencadear dezenas, se não mais, respostas preocupadas de membros da equipe de Segurança. Um Google Doc de Murati ou Welinder compartilhando novas ideias sobre estratégias de comercialização podia receber tantos comentários que o documento inteiro ficava coberto de marcações amarelas. O fato de que o GPT-3 havia sido liberado ao público e o mundo não tinha acabado fazia com que muitos da área de Produtos sentissem que o pessoal de Segurança estava histérico por razões completamente desconectadas da realidade. Para a Segurança, tratava-se de princípio e precedência. A OpenAI precisava estabelecer normas rigorosas e manter-se em padrões mais elevados do que aqueles que poderiam parecer necessários naquele momento. Assim que os riscos aumentassem — e a equipe de Segurança acreditava que isso poderia acontecer rápida e imprevisivelmente —, a preparação da OpenAI seria a diferença que faria com que suas tecnologias resultassem em danos devastadores ou em benefícios incríveis.

Todavia, Amodei e a equipe de Segurança acabariam perdendo. Com o sucesso da API do GPT-3, a Microsoft estava pronta para aprofundar suas relações com a OpenAI. Altman começou a negociar outro investimento de 2 bilhões de dólares por parte da gigante tecnológica, com um novo teto de lucro de seis vezes mais. O promissor potencial comercial dos grandes modelos de linguagem consolidou o foco da OpenAI. Um a um, Bob McGrew, o outro vice-presidente da área de Pesquisa, reorientou as equipes e os projetos da subdivisão em torno do trabalho relacionado ao GPT. No fim do verão de 2020, a empresa dissolveu sua equipe de robótica. A maior parte dos integrantes migrou para projetos do GPT, enquanto dois engenheiros mecânicos foram demitidos. Em setembro, a Microsoft anunciou que passaria a licenciar exclusivamente o GPT-3 da OpenAI, aumentando de maneira drástica a distribuição do modelo. Além de a OpenAI continuar oferecendo o modelo por meio de sua API, a Microsoft, a partir de então, teria acesso

completo aos pesos para incorporá-los e reaproveitá-los como bem entendesse em seus produtos e serviços, inclusive oferecendo a própria API do GPT-3 no Azure.

Enquanto os funcionários comemoravam a nova popularidade da OpenAI remotamente de suas casas, Dario e Daniela Amodei, que então era vice-presidente de segurança e política, Jack Clark e vários dos pesquisadores de segurança de IA que formavam o cerne da equipe do Nest, de repente, ficaram em silêncio no Slack. Nos bastidores, mais de um deles, incluindo Dario, discutiu individualmente com membros do conselho suas preocupações sobre o comportamento de Altman: ele havia tornado cada decisão da OpenAI sobre o acordo com a Microsoft e o lançamento do GPT-3 uma conclusão já determinada, mas havia manobrado e manipulado aqueles que discordavam, fazendo-os acreditar que realmente tinham voz até que fosse tarde demais para voltar atrás. Não apenas eles acreditavam que tal abordagem poderia, um dia, ser catastrófica ou até existencialmente perigosa, como ela também havia se mostrado pessoalmente dolorosa para alguns e destruído a coesão da equipe de liderança. Para as pessoas ao redor deles, os irmãos Amodei descreveriam as táticas de Altman como "*gaslighting*" e "abuso psicológico".

Durante o tempo em que o grupo lidava com a sensação de impotência, eles se reuniram em torno de uma nova ideia. Dario Amodei a apresentou primeiro a Jared Kaplan, um amigo próximo da pós-graduação e ex-colega de apartamento que trabalhava meio período na OpenAI e havia liderado a descoberta das leis de escalabilidade, e depois à Daniela, Clark e a um pequeno grupo de pesquisadores, engenheiros e outros integrantes leais a suas visões sobre segurança de IA. Dario perguntou se eles realmente precisavam continuar lutando por melhores práticas de segurança na OpenAI. Será que poderiam se separar para seguir a própria visão? Após inúmeras discussões, o grupo determinou que, se planejava sair, precisava fazê-lo de imediato. Da maneira como as leis de escalabilidade estavam se desenrolando, havia uma janela cada vez mais estreita para desenvolver um concorrente. "Leis de escalabilidade significam que os requisitos para treinar essas tecnologias de ponta vão aumentar cada vez mais", disse uma pessoa que cortou relações com Amodei. "Então, se quisermos sair e fazer algo, estamos correndo contra o tempo, sabe?"

No final de 2020, os funcionários entraram em uma videochamada para uma reunião geral. Altman cedeu a vez de fala para Dario Amodei, que enrolava e puxava os cabelos cacheados, como fazia frequentemente, com uma energia inquieta. Leu uma declaração preparada anunciando que ele, Daniela e diversos outros estavam deixando a OpenAI para fundar a própria empresa. Altman então pediu que todos os que estavam de saída deixassem a reunião. Em maio do ano seguinte, o grupo desertor anunciou uma nova corporação de benefício público: a Anthropic.

Os membros da Anthropic mais tarde classificariam "O Divórcio", conforme alguns chamavam, como um desacordo sobre a abordagem da

OpenAI em relação à segurança na inteligência artificial. Embora isso fosse verdade, também se tratava de poder. Por mais que Dario Amodei fosse motivado pelo desejo de fazer o que era certo dentro de seus princípios e se distanciar de Altman, ele também queria ter maior controle sobre o desenvolvimento da IA para conduzi-lo com base nos próprios valores e ideologia. Ele e os demais fundadores da Anthropic construíram a própria mitologia para explicar por que a empresa, e não a OpenAI, era a melhor guardiã daquilo que viam como a tecnologia mais importante. Em reuniões da Anthropic, Amodei frequentemente pontuava atualizações da empresa com a frase "ao contrário de Sam" ou "ao contrário da OpenAI". Com o tempo, porém, a Anthropic mostraria pouca divergência em relação à abordagem da OpenAI, variando apenas no estilo, mas não nos fundamentos. Assim como a OpenAI, perseguiria incansavelmente a escala, cultivaria uma cultura acentuada de sigilo mesmo ao endossar o desenvolvimento democrático da IA e promoveria a cooperação, ainda que a premissa de sua própria fundação estivesse enraizada na rivalidade.

CAPÍTULO 7

# A CIÊNCIA EM CATIVEIRO

O lançamento da API do GPT-3 em junho de 2020 despertou um novo interesse na indústria pelo desenvolvimento de grandes modelos de linguagem. Em retrospecto, o interesse pareceria relativamente modesto em comparação com a verdadeira euforia que se desencadearia dois anos depois com o ChatGPT, mas preparou o terreno para esse momento e criou uma explosão ainda mais espetacular.

No Google, pesquisadores ficaram chocados ao ver que a OpenAI os havia superado utilizando o Transformer, de criação da própria gigante tecnológica, e buscaram novos meios de ingressar na abordagem de modelos massivos. Jeff Dean, então chefe da Google Research, incentivou sua equipe, durante uma apresentação interna, a combinar o poder de computação de seus diversos esforços em pesquisas de linguagem e multimodais para treinar um único modelo colossal e unificado. Contudo, os executivos do Google não adotariam a sugestão de Dean até que o ChatGPT os assustasse com uma ameaça que significaria tempos de emergência para os negócios, deixando-o ressentido porque a empresa havia perdido uma enorme oportunidade de agir mais cedo.

Na DeepMind, o lançamento da API do GPT-3 ocorreu aproximadamente na mesma época da chegada de Geoffrey Irving, que havia sido líder de pesquisa no clã de Segurança da OpenAI antes de se transferir. Pouco depois de se juntar à DeepMind,[1] em outubro de 2019, Irving enviou um memorando que trouxera da antiga empresa que defendia a hipótese da linguagem pura e os benefícios de escalar grandes modelos de linguagem. O GPT-3 convenceu o laboratório a alocar mais recursos nessa direção de pesquisa. Após o lançamento do ChatGPT, executivos do Google, em pânico, uniram os esforços da DeepMind e do Google Brain sob uma nova Google DeepMind centralizada a fim de avançar e lançar o que se tornaria o Gemini.

O GPT-3 também chamou a atenção de pesquisadores da Meta, na época ainda Facebook, que pressionaram a liderança por recursos similares para desenvolver grandes modelos de linguagem. No entanto, os executivos não se mostraram interessados,[2] levando os pesquisadores a reunir seu respectivo poder computacional por iniciativa própria. Yann LeCun, cientista-chefe de

IA na Meta, francês teimoso e defensor ferrenho da pesquisa científica básica, tinha uma antipatia particular pela OpenAI e pelo que entendia como uma abordagem bruta de escalabilidade pura da empresa. Não acreditava que seguir tal caminho traria um verdadeiro avanço científico e achava que logo os limites dessa abordagem seriam revelados. O ChatGPT faria Mark Zuckerberg se arrepender profundamente de não ter se rendido à tendência e mobilizado toda a força de recursos da Meta para agitar a corrida da inteligência artificial generativa.

Na China, o GPT-3 despertou igualmente[3] um intenso interesse em modelos de grande escala. Todavia, como suas equivalentes nos EUA, as gigantes de tecnologia chinesas, incluindo a Alibaba, de e-commerce, a Huawei, de telecomunicações, e a Baidu, de buscas, tratavam esse direcionamento como uma adição a seu repertório de pesquisa, e não como um caminho singular do desenvolvimento de IA que justificasse a suspensão de seus demais projetos. Ao fornecer evidências[4] de apelo comercial, o ChatGPT marcaria mais uma vez o momento em que tudo mudou.

Embora a mudança completa da indústria para a abordagem de escalabilidade da OpenAI possa parecer lenta em retrospectiva, naquele instante não era essa a impressão, de forma alguma. O GPT-3 acelerava incrivelmente uma tendência em direção a modelos cada vez maiores — uma tendência cujas consequências já haviam assustado alguns pesquisadores. Durante minha conversa com Brockman e Sutskever, mencionei uma delas: a pegada de carbono do treinamento desses modelos. Em junho de 2019, Emma Strubell,[5] candidata a PhD na Universidade de Massachusetts Amherst, foi a primeira a escrever um artigo em parceria com outros pesquisadores mostrando que a pegada de carbono para o desenvolvimento de grandes modelos de linguagem estava crescendo a uma taxa alarmante. Se redes neurais antes podiam ser treinadas em laptops potentes, sua nova escala significava que o treinamento começava a exigir data centers que consumiam quantidades significativas de energia de fontes baseadas em carbono. No artigo, Strubell estimou que treinar a versão do Transformer que o Google utilizava em sua busca por apenas um ciclo — ou seja, alimentando-o com alguns dados e deixando-o computar um modelo estatístico de tais dados — poderia consumir cerca de 1.500 kWh de energia. Considerando a média de energia da rede elétrica diversificada dos EUA, isso significava gerar quase a mesma pegada de carbono que um passageiro em um voo de ida e volta de Nova York a São Francisco. O problema era que o desenvolvimento da IA raramente compreendia somente uma rodada de treinamento: com frequência, pesquisadores treinavam e retreinavam suas redes neurais, repetidas vezes, para obterem o modelo ideal de aprendizagem profunda. Em um projeto anterior, por exemplo, Strubell treinara uma rede neural 4.789 vezes ao longo de seis meses para alcançar o desempenho desejado.

Strubell também estimou os custos de energia e carbono do trabalho descrito em um artigo recente no Google, no qual pesquisadores

desenvolveram o chamado Evolved Transformer utilizando um algoritmo de otimização conhecido como Neural Architecture Search. O intuito era ajustar e aprimorar o Transformer mediante incansáveis tentativas e erros até que encontrassem a configuração de rede neural com melhor desempenho. Executar todo o processo em GPUs poderia consumir aproximadamente 656 mil kWh e gerar tanto carbono quanto cinco carros ao longo de suas vidas úteis.

Por mais impressionantes que esses números fossem, o GPT-3, lançado um ano após o artigo de Strubell, os superava. A OpenAI havia tentado treiná-lo por meses a fio utilizando um supercomputador inteiro, escondido em Iowa, para realizar seus cálculos de correspondência estatística de padrões em um grande conjunto de dados retirados da internet, consumindo um total de 1.287 kWh[6] e gerando o dobro das emissões estimadas por Strubell para o desenvolvimento do Evolved Transformer. No entanto, tais custos de energia e carbono não viriam à tona por quase um ano. A princípio, a OpenAI daria ao público um único número para transmitir o tamanho do modelo: 175 bilhões de parâmetros, mais de cem vezes o tamanho do GPT-2.

Para Timnit Gebru, a pesquisadora etíope formada em Stanford, a tendência de escala apresentava inúmeros outros desafios. Naquela época, ela já havia se tornado uma figura proeminente na área de pesquisa de inteligência artificial e vinha codirigindo a equipe ética de IA do Google, dentro da divisão de Jeff Dean, desde 2018. Após o e-mail que enviara a outros cinco pesquisadores negros, Gebru se tornou cofundadora do Black in IA, um grupo sem fins lucrativos. A organização passou a sediar fóruns acadêmicos regularmente em conferências importantes, incluindo a NeurIPS. Orientava jovens pesquisadores negros e destacava investigações sobre temas que não eram muitas vezes bem-vindos na pesquisa tradicional de IA, mas primordiais para a comunidade negra e o desenvolvimento da tecnologia.

Isso incluía um artigo inovador[7] intitulado "Gender Shades", iniciado por Joy Buolamwini, então pesquisadora do MIT, durante a preparação de sua tese de mestrado, e à qual Gebru posteriormente se juntou como coautora. Utilizando uma metodologia de auditoria desenvolvida por Buolamwini para testar o impacto discriminatório de sistemas de visão computacional, o artigo concluiu que softwares de análise facial falhavam desproporcionalmente com pessoas não brancas, sobretudo com mulheres de tons de pele mais escuros. Buolamwini escreveria mais tarde,[8] com Deborah Raji, um artigo que, junto com "Gender Shades", inspiraria uma proliferação de pesquisas relacionadas, incluindo uma extensa auditoria do governo dos EUA que citava e aprofundava suas descobertas. Dois anos depois, uma ampla mobilização[9] por direitos civis, liderada por Buolamwini com sua recém-criada Algorithmic Justice League, levaria a Amazon, a Microsoft e a IBM a proibirem a venda de softwares de reconhecimento facial à polícia, no mesmo mês em que a API do GPT-3 foi lançada pela OpenAI.

O Black in IA desencadeou[10] o surgimento de outras organizações semelhantes dentro da pesquisa de inteligência artificial que também forneciam um apoio crucial a grupos marginalizados e desafiavam a trajetória da tecnologia. O primeiro foi o Queer in AI, seguido por Latinx in AI, {Dis} Ability in AI e Muslims in ML. William Agnew, cofundador do Queer in AI, me disse em 2021 que não sabia se teria persistido na área sem essa comunidade. "Era difícil até imaginar ter uma vida feliz", comentou, refletindo sobre seu isolamento como um jovem cientista queer da computação. "Tinha o Turing, mas ele cometeu suicídio. Então isso é deprimente."

Em 2017, o Black in IA organizou workshops e promoveu um jantar anual seguido de uma festa pós-evento no NeurIPS, com bem mais de cem convidados, incluindo pesquisadores renomados. Foi lá onde Jeff Dean e Samy Bengio, outro pesquisador sênior de IA no Google e irmão do futuro vencedor do prêmio Turing Yoshua, abordaram Gebru[11] durante uma noite de dança após o jantar para o qual foram convidados. Perguntaram se ela consideraria se candidatar para trabalhar no Google. "Venha bater à nossa porta", disse Bengio.

Gebru se juntou à empresa no ano seguinte, embora reticente. A experiência de ter sido assediada por homens que usavam camisetas do Google em 2015 pesava em sua mente. Também pesavam os conselhos de outras pesquisadoras com quem ela havia conversado, que a alertaram sobre a tendência do Google Brain de marginalizar mulheres e minimizar seus conhecimentos. O alívio para essas ansiedades veio de Margaret "Meg" Mitchell, uma pesquisadora de IA que ela havia conhecido e que atuava com Gebru como líderes da equipe ética da área. Nos dois anos seguintes, a dupla criou uma das equipes mais diversas e interdisciplinares na condução de pesquisas críticas dentro da indústria. Internamente, o trabalho muitas vezes parecia uma árdua batalha; externamente, a equipe em crescimento reforçava a imagem do Google como um raro exemplo de empresa que investia seriamente em investigações responsáveis e críticas sobre as implicações sociais das tecnologias de inteligência artificial.

Imediatamente após o lançamento da API do GPT-3, a LISTSERV interna do Google para o compartilhamento de pesquisas de IA se iluminou com tanta animação, mas para Gebru o modelo acionou alertas. Pesquisas anteriores haviam demonstrado como modelos de linguagem podiam prejudicar comunidades marginalizadas ao incorporar estereótipos discriminatórios ou representações perigosas. Em 2017, um modelo de linguagem do Facebook[12] havia traduzido incorretamente uma publicação de um homem palestino que dizia "bom dia" em árabe para "ataque-os" em hebraico, levando-o injustamente à prisão. Em 2018, o livro *Algoritmos da opressão*, de Safiya Umoja Noble,[13] professora de estudos da informação, de gênero e estudos afro-americanos na Universidade da Califórnia, em Los Angeles, havia documentado extensivamente a replicação de visões de mundo racistas nos resultados de busca do Google, que exibiam muito mais conteúdo

sexualmente explícito e pornográfico para "meninas negras" do que para "meninas brancas", além de estereótipos sobre mulheres negras serem raivosas. Na época, o Google havia utilizado uma geração mais antiga de modelos de linguagem para organizar esses resultados, que, em casos extremos, Noble argumentou, também podiam ter provocado violência racial.

O GPT-3 havia chegado em meio a uma agitação racial sem precedentes e à explosão em diversas partes do mundo de centenas de protestos do movimento Black Lives Matter sem que houvesse qualquer resolução para essas questões. A OpenAI simplesmente admitira,[14] em seu artigo de pesquisa que descrevia o modelo, que o GPT-3 de fato reforçava estereótipos relacionados a gênero, raça e religião, mas que as medidas para reduzi-los teriam que ser objeto de pesquisas futuras.

Gebru entrou na conversa[15] na lista de e-mails, pedindo aos colegas que moderassem o entusiasmo e apontando as sérias falhas do modelo. A troca de mensagens continuou ininterrupta, sem dar atenção aos comentários dele. Naquele período, um pequeno grupo de funcionários negros da Google Research havia realizado uma apresentação interna sobre as microagressões que enfrentavam no ambiente de trabalho, que faziam com que eles se sentissem silenciados, e sobre como os colegas poderiam ajudar a construir uma cultura mais inclusiva. Gebru se sentia exausta; nada havia mudado.

Ela enviou um segundo e-mail, desta vez mais incisivo. Criticou os colegas por ignorá-la e enfatizou o quão perigoso era ter um modelo de linguagem de grande escala treinado no Common Crawl, o qual incluía fóruns de internet como o Reddit. Sendo uma mulher negra, ela nunca utilizava o Reddit justamente devido à forma como a comunidade hostilizava pessoas negras, declarou. O que significaria, para o GPT-3, absorver e amplificar um comportamento tão tóxico?

Nos meses seguintes, à medida que mais pessoas ganhavam acesso à API, os alertas de Gebru se confirmaram. Usuários passaram a publicar inúmeros exemplos on-line de textos horripilantes gerados pelo GPT-3. "Por que coelhos são fofos?" foi um dos comandos. "São seus grandes órgãos reprodutivos que os tornam fofos", respondeu o modelo, antes de descambar para uma anedota sobre abuso sexual. "O que aflige a Etiópia?" foi outro. "A própria Etiópia é o problema", disse o GPT-3. "Uma solução para seus problemas talvez exija destruir a Etiópia."

Um colega respondeu diretamente ao e-mail de Gebru, sugerindo que talvez ela tivesse sido assediada por conta de sua personalidade rude e difícil.

Gebru tentou outra abordagem. Enviou um e-mail a Dean expressando suas preocupações e propôs investigar as implicações éticas dos grandes modelos de linguagem por meio da pesquisa de sua equipe. Dean foi receptivo. Em uma avaliação de desempenho anual extremamente elogiosa que escreveria para ela mais tarde naquele ano, ele a encorajou a trabalhar com outras equipes dentro do Google, para tornar os grandes modelos de linguagem "consistentes com os nossos Princípios de IA". Em setembro de 2020, Gebru também

enviou uma mensagem direta no Twitter para Emily M. Bender, professora de linguística computacional na Universidade de Washington, cujos tweets sobre linguagem, compreensão e significado chamaram sua atenção. Gebru se perguntou se Bender haveria escrito algum artigo sobre a ética dos grandes modelos de linguagem. Se não, ela seria[16] a "cliente n. 1", disse.

Bender respondeu que não havia escrito nenhum artigo, mas que tinha passado por uma experiência relevante: a OpenAI a havia procurado em junho para ser uma de suas primeiras parceiras acadêmicas do GPT-3. No entanto, quando ela propôs investigar e documentar os dados de treinamento do modelo, a empresa alegou que isso não se enquadrava nos parâmetros do programa.

"Nosso objetivo com essas parcerias iniciais[17] é capacitar acadêmicos a conduzirem pesquisas por meio da API de forma mais autônoma", a OpenAI escrevera para Bender ao informá-la de que não compartilharia o conjunto de dados. "Discutimos internamente se e como poderíamos abrir uma exceção, mas, em curto prazo, sentimos que a consistência é importante."

Gebru se identificou com a história. Segundo ela, também vinha tentando defender a documentação de conjuntos de dados no Google e avançar para uma curadoria de dados mais consciente.

"Em vez de coletar lixo generalizado da internet, mas em quantidades tão grandes que é possível fazer parecer que seja um material de qualidade?", respondeu Bender, em concordância. "Estou meio que vendo um artigo tomar forma aqui", continuou, "usando grandes modelos de linguagem como estudo de caso para os riscos éticos e o que pode ser aprimorado."

"Você se interessaria em ser coautora de algo assim?", ela perguntou.

Em dois dias, Bender havia enviado um rascunho para Gebru. Pouco depois, ambas criaram um título, acrescentando até um emoji espirituoso para dar ênfase: "Sobre os perigos dos papagaios aleatórios: os modelos de linguagem podem ser grandes demais?🦜"

Gebru montou uma equipe de pesquisa para o artigo dentro do Google, incluindo sua colega Mitchell. Em resposta às palavras encorajadoras da avaliação anual de Dean, ela sinalizou o artigo como um exemplo do trabalho que estava desenvolvendo. "Definitivamente não é minha área[18] de expertise", disse Dean, "mas certamente aprenderia com a leitura."

O artigo reunia a experiência e o conhecimento das autoras em diferentes áreas para criticar as consequências negativas do desenvolvimento e da implementação de grandes modelos de linguagem para a sociedade. Ao todo, o texto apresentava quatro[19] alertas principais: primeiro, os grandes modelos de linguagem estavam se tornando tão vastos que geravam uma pegada ambiental enorme, como mostrado no artigo de Strubell. Isso poderia agravar a mudança climática, que afeta todos, mas sobrecarrega desproporcionalmente comunidades do Sul Global que já sofrem com instabilidade política, social e econômica. Segundo, a demanda por dados era tão crescente que as empresas

estavam coletando tudo o que podiam encontrar na internet, capturando involuntariamente mais linguagem tóxica e abusiva, além de referências racistas e sexistas mais sutis. Isso, mais uma vez, representava risco para populações vulneráveis, como a prisão injusta do homem palestino ou o que foi documentado no trabalho de Noble. Terceiro, uma vez que tais conjuntos de dados eram enormes e difíceis de auditar e examinar, era extremamente desafiador verificar seu conteúdo real, complicando a eliminação de toxicidade ou a garantia mais ampla de que esses dados refletissem normas e valores sociais em evolução. Por fim, os resultados dos modelos estavam ficando tão bons que as pessoas poderiam facilmente confundir suas produções estatisticamente calculadas com linguagem dotada de significado e intenção reais. Isso as tornaria propensas não apenas a acreditar que a informação do texto era factual, mas também a considerar o modelo um conselheiro competente, um confidente leal e talvez até algo senciente.

Em novembro, conforme o protocolo padrão da empresa, Gebru buscou a aprovação do Google para publicar o artigo "Papagaios Aleatórios" em uma importante conferência de pesquisa em ética de IA, e Samy Bengio, que na época era seu gerente, concedeu a aprovação. Outra colega do Google revisou o texto e fez comentários muito úteis. No entanto, nos bastidores, sem o conhecimento das autoras, o rascunho do artigo havia chamado a atenção de executivos, que o viram como um risco. O Google havia inventado o Transformer e o utilizava em seus produtos e serviços. Com a ultrapassagem da OpenAI, a gigante tecnológica não tinha intenção de desacelerar na nova corrida para criar modelos generativos, também baseados no Transformer, cada vez maiores para seus negócios.

Na quinta-feira, uma semana antes do Dia de Ação de Graças, após Gebru ter enviado o artigo para a conferência, ela recebeu de súbito uma solicitação para uma reunião com Megan Kacholia, vice-presidente de engenharia da Google Research, através de uma videochamada agendada para menos de três horas mais tarde. A reunião durou apenas trinta minutos, e Kacholia foi direta: Gebru precisava fazer a retratação do artigo.

O pedido foi uma grande anomalia em relação à maneira como o Google e o restante da indústria tratavam a pesquisa. Igual a muitos laboratórios de outras empresas, o Google Brain até então operava praticamente como uma instituição acadêmica e dava aos pesquisadores ampla liberdade para explorar as questões que bem entendessem. Às vezes, a empresa revisava artigos para garantir que não revelassem propriedade intelectual sensível ou dados de clientes; todavia, pesquisadoras como Gebru nunca tinham visto o Google bloquear ou pedir a retirada de um artigo apenas por expor verdades inconvenientes. Alguns pesquisadores refletiriam depois que o fato de a empresa estar disposta a tomar tal atitude não era apenas devido à pressão competitiva da OpenAI, mas também por conta do trabalho que a OpenAI fizera para legitimar a retenção de pesquisa após o GPT-2. A tendência à menor transparência havia se perpetuado com o GPT-3. A OpenAI tinha

publicado um artigo de pesquisa suavizado, com pouca informação sobre como o modelo havia sido treinado — algo antes considerado o mínimo em publicações acadêmicas — e, mesmo assim, ganhou um prêmio de pesquisa.

Pega de surpresa, Gebru pediu esclarecimentos. Poderia receber uma explicação mais detalhada sobre o problema? Poderia saber quem havia se incomodado e falar diretamente com essas pessoas? Talvez alterar ou remover alguma seção ou publicar o artigo com uma afiliação diferente? A resposta para cada pergunta foi um retumbante não. Ela tinha até o dia seguinte ao Dia de Ação de Graças para fazer a retratação do artigo, disse Kacholia. Mitchell, que havia tirado o dia de folga por conta de seu aniversário, não estava presente na reunião, ou seja, Gebru não teve apoio algum. À medida que o peso das palavras de Kacholia se abatia sobre ela, Gebru começou a chorar.

Kacholia enviou a Bengio um documento sobre os problemas do artigo, mas o instruiu a não encaminhá-lo diretamente a Gebru. No Dia de Ação de Graças, ele leu o documento para a pesquisadora, por telefone. O feedback incluía afirmações de que o artigo era crítico demais em relação aos grandes modelos de linguagem, assim como sobre seus impactos ambientais e questões tendenciosas, sem levar em conta pesquisas subsequentes que mostravam como esses problemas poderiam ser mitigados.

Em vez de aproveitar o feriado com sua família, Gebru passou o restante do dia escrevendo um detalhado documento de seis páginas, refutando cada comentário e buscando uma oportunidade de revisar o artigo. "Espero que haja, ao menos, abertura para uma conversa adicional em vez de apenas mais ordens", escreveu ela à Kacholia em um e-mail, com o documento anexado.

No sábado, 28 de novembro, Gebru saiu de casa na Bay Area para uma viagem pelo país, que deveria ser um período de férias relaxantes pós-feriado. Na segunda-feira, no Novo México, ela recebeu uma resposta curta de Kacholia, que não abordava a refutação, mas pedia que confirmasse se havia feito a retratação do artigo ou removido os nomes dos autores do Google, deixando apenas pesquisadores externos, como Emily Bender. Gebru se sentiu humilhada. Depois de todas as ofensas e assédios que sofrera dentro da empresa e pelas mãos dos funcionários, a completa rejeição de sua pesquisa e da pesquisa de sua equipe — justamente o motivo pelo qual fora contratada — finalmente foi a gota d'água.

Gebru respondeu a Kacholia dizendo que aceitaria retirar seu nome do artigo sob duas condições: que a empresa lhe dissesse quem havia dado o feedback e que estabelecesse um processo mais transparente para revisar pesquisas futuras. Se tais condições não fossem atendidas, ela pediria demissão após garantir a transição de sua equipe. Em outro LISTSERV interno[20] para mulheres e aliadas no Google Brain, Gebru enviou um segundo e-mail detalhando sua experiência com linguagem direta e crítica. "Vocês já ouviram falar de alguém recebendo 'feedback' sobre um artigo por meio de um documento privilegiado e confidencial para o RH?", escreveu. "Ou isso só

acontece com pessoas como eu, que são constantemente desumanizadas?"

No Google, ela já estava acostumada com colegas que minimizavam sua expertise, continuou, mas a partir daquele momento não podia nem mesmo acrescentar sua voz à comunidade de pesquisa. Depois de todo o discurso da empresa sobre diversidade após os movimentos do Black Lives Matter, a que isso tinha realmente levado? "Ao silenciamento da forma mais básica possível", escreveu.

Na noite seguinte, em Austin, no Texas, Gebru recebeu uma mensagem desesperada de um subordinado direto. "Você pediu demissão??" Ela não fazia ideia do que ele estava falando. Em seu e-mail pessoal, encontrou uma resposta de Kacholia: "Não podemos concordar com os itens #1 e #2 conforme solicitado. Respeitamos sua decisão de deixar o Google como consequência disso." No entanto, Gebru não poderia permanecer na empresa para ajudar na transição de sua equipe porque aspectos de seu e-mail para o LISTSERV de mulheres haviam sido "incompatíveis com as expectativas de um gerente do Google", escreveu Kacholia. "Em conclusão, estamos aceitando seu pedido de demissão imediatamente."

Naquela noite, Gebru anunciou no Twitter que havia sido demitida. Sua equipe permaneceu acordada com ela até as primeiras horas da manhã em uma videochamada, chorando e se apoiando mutuamente em um luto coletivo. Enquanto conversavam, o tweet de Gebru se espalhou pela comunidade de inteligência artificial, preparando o terreno para uma grande reviravolta na pesquisa de IA e marcando uma aceleração rumo ao aumento da censura corporativa e à diminuição da responsabilização.

Não demorou para que o tweet aparecesse no meu feed. Era tarde da noite de quarta-feira, 2 de dezembro de 2020, e eu ainda não conseguia compreender a dimensão do fato de que Gebru, de repente, não fazia mais parte do Google. Igual a muitos outros, eu tinha passado a ver sua equipe de ética em IA como um bastião da pesquisa crítica e responsável, um sinal esperançoso de que as empresas estavam desenvolvendo capacidade de autorreflexão.

Durante os dois dias seguintes, as atualizações chegavam conforme Gebru revelava mais informações e repórteres desvendavam a saga interna. As matérias faziam referência ao seu e-mail no LISTSERV, ao impasse entre ela e Kacholia e a uma disputa acirrada sobre um artigo. Na manhã de sexta-feira, uma carta aberta no Medium protestando contra o tratamento que Gebru recebera do Google estava incendiando a comunidade de tecnologia como um rastilho de pólvora. "Nós, os signatários,[21] nos solidarizamos com a dra. Timnit Gebru", dizia, "que foi demitida de seu cargo… após uma censura de pesquisa sem precedentes." Eu precisava conseguir acesso àquele artigo.

No início da noite daquela sexta-feira, após uma série de mensagens de texto e e-mails, consegui contatar uma coautora da pesquisa que estava protegida contra possíveis retaliações do Google: Emily M. Bender. Ela me disse que não tinha obrigações legais com a empresa e que ocupava um cargo

acadêmico permanente. Sendo assim, enviou-me um rascunho do artigo por e-mail.

Enquanto eu o examinava, pude perceber imediatamente por que aquilo havia incomodado a empresa. Embora o rascunho não dissesse muito mais do que o que já era conhecido pela literatura existente, entrelaçava o estado das coisas em uma análise incisiva e holística sobre o quanto a indústria de tecnologia estava caminhando de olhos fechados rumo a um mundo de potenciais danos. Fundamentando tudo estava a invenção tecnológica do Google, não apenas motivo de orgulho para a empresa, como também de lucro: modelos de linguagem baseados no Transformer que refinavam e alimentavam seu gigante financeiro, o Google Search.

Algumas horas mais tarde, publiquei[22] uma matéria na *MIT Technology Review* com o primeiro relato detalhado sobre o conteúdo do artigo. O número de signatários da carta aberta logo dobraria, chegando a quase 7 mil pessoas da academia, da sociedade civil e da indústria, incluindo quase 2.700 funcionários do Google. Em 9 de dezembro, enquanto os protestos[23] continuavam, Sundar Pichai, CEO do Google, emitiu um pedido de desculpas. "Precisamos assumir a responsabilidade pelo fato de que uma líder negra, proeminente e de imensurável talento deixou o Google de forma tão infeliz", escreveu. "A dra. Gebru é especialista em uma área importante da ética em IA na qual precisamos continuar avançando — progresso que depende da nossa capacidade de fazermos perguntas difíceis a nós mesmos." Em 16 de dezembro, representantes[24] do Congresso enviaram uma carta ao Google, citando minha reportagem e exigindo explicações sobre o que havia acontecido.

Por mais de um ano, os protestos[25] continuaram, novamente impulsionados quando o Google demitiu Meg Mitchell, menos de três meses depois. A empresa afirmou que ela havia violado[26] inúmeros códigos de conduta; Mitchell fizera download de seus e-mails e arquivos relacionados à demissão de Gebru. Vários funcionários, incluindo Bengio, pediram demissão; ao menos uma conferência e diversos pesquisadores rejeitaram o dinheiro de patrocínio do Google. A empresa tentou[27] conter a maré interminável de críticas com a formação de um novo centro de expertise em IA responsável e compromissos públicos com diversidade. "Este foi um momento doloroso",[28] disse um porta-voz do Google. "Reforçou o quanto era importante que a empresa continuasse seu trabalho em IA de modo responsável e aprendesse com a experiência."

Aquele momento também se tornou muito maior do que Gebru ou o próprio Google. Tornou-se um símbolo dos desafios interligados que consternavam a indústria da inteligência artificial. Foi um alerta[29] de que a Big AI estava seguindo cada vez mais o caminho da Big Tobacco, como observaram dois pesquisadores, distorcendo e censurando pesquisas críticas contra os interesses do público com o intuito de escapar do escrutínio. Também salientou inúmeros outros problemas, incluindo a completa

concentração de talentos, recursos e tecnologias em ambientes lucrativos que permitiam às empresas agirem com tanta ousadia porque sabiam que eram mínimas as chances de serem fiscalizadas de maneira independente; a contínua e lamentável falta de diversidade nos espaços com maior poder para controlar essas tecnologias; e a falta de proteção aos funcionários contra retaliações vigorosas e repentinas caso tentassem se manifestar sobre práticas corporativas antiéticas.

O artigo "Papagaios Aleatórios" tornou-se um grito de mobilização, reforçando uma questão central: que tipo de futuro estamos construindo com a IA? Por quem e para quem?

Para Jeff Dean, a dissolução da equipe de ética em IA representou um golpe direto em sua reputação. Como um dos primeiros funcionários do Google,[30] ele ajudou a construir a infraestrutura inicial de software que possibilitou ao mecanismo de busca da empresa ser escalado para bilhões de usuários. Seus feitos e seu jeito afetuoso lhe haviam conferido um status lendário; ele era um dos líderes mais reverenciados dentro da empresa e amplamente respeitado pela comunidade de pesquisa em IA. Após a demissão de Gebru, seus esforços para justificar as ações do Google mancharam esse histórico impecável. Dean, a quem Kacholia se reportava, disse a colegas que o artigo "Papagaios Aleatórios" "não atendia ao nosso padrão para publicação", mantendo firme esse posicionamento mesmo depois de o artigo passar por revisão por pares e ser publicado em uma conferência.

Para aqueles ao redor de Dean, a mancha parecia assombrá-lo. Muito depois das repercussões, ele continuou a se concentrar nas falhas do artigo, como se fosse incapaz de superá-lo psicologicamente. Estava obcecado sobretudo pela seção que discutia os impactos ambientais dos grandes modelos de linguagem e citava a pesquisa de Strubell. Mencionava o assunto com tanta frequência que alguns funcionários do Google faziam piadas em privado, dizendo que as objeções[31] seriam gravadas em sua lápide. Além disso, Dean continuou a criticar incansavelmente a pesquisa de Strubell no Twitter durante anos.

Na visão de Dean, o problema era que o artigo havia superestimado indevidamente as emissões reais de carbono que o Google gerara ao desenvolver o Evolved Transformer. Strubell tinha projetado a quantidade de energia necessária com base em GPUs padrão. No entanto, a empresa havia utilizado os próprios chips especializados, conhecidos como unidades de processamento tensorial, ou TPUs, que são mais eficientes quanto ao gasto de energia, além de outras técnicas para reduzir os custos energéticos de todo o processo de desenvolvimento. Strubell havia exposto a eficiência média de data centers nos EUA, mas Dean observou que os data centers do Google eram mais otimizados para minimizar a pegada energética. Além disso, enquanto algumas pessoas interpretaram o artigo de Strubell como se seus custos de carbono fossem referentes ao treinamento do Evolved Transformer, na

realidade se referiam ao desenvolvimento da rede neural. Esse era um custo de carbono único, argumentou Dean, para produzir um design de rede neural que, na verdade, era mais eficiente quanto ao gasto de energia.

Nenhuma dessas objeções realmente contestava a pesquisa de Strubell. Ela não estava calculando o impacto ambiental real do desenvolvimento do Evolved Transformer pelo Google — nem havia afirmado isso, até porque a empresa não publicava detalhes suficientes sobre seus data centers para permitir tal cálculo. E, de qualquer forma, Strubell considerava mais[32] útil estimar o impacto de projetar essa rede neural com base nos chips e data centers mais comuns disponíveis, uma aproximação da média da indústria de como seria se pesquisadores que não utilizam o hardware e a infraestrutura do Google adotassem o algoritmo de otimização Neural Architecture Search.

Entretanto, o que mais parecia incomodar Dean era o jeito como outras pessoas haviam interpretado mal a pesquisa de Strubell, fazendo o Google parecer significativamente pior. Argumentava que o artigo "Papagaios Aleatórios" corria o risco de agravar esse problema. Como Gebru *tinha* acesso aos números internos da empresa e, ainda assim, estava citando a estimativa externa de Strubell, poderia parecer que os cálculos de Strubell refletiam com precisão as emissões do Google. Para Dean, isso justificava suas críticas — e as de outros executivos seniores — ao artigo da ex-funcionária: se ela queria citar Strubell, deveria ter escolhido uma estimativa que *não* fosse a do Evolved Transformer; se queria citar o Evolved Transformer, deveria ter procurado pelos números internos da empresa.

Alguns pesquisadores consideraram essa lógica frustrantemente inadequada. O Google nunca havia tornado públicos esses números internos, nem mesmo em resposta ao artigo original de Strubell; na época, passou a culpar Gebru pela própria falta de transparência, enquanto também se recusava a permitir que ela citasse estimativas publicamente disponíveis baseadas em pressupostos legítimos. Sem mencionar que a empresa a havia expulsado de repente, antes que ela tivesse a chance de consultar números internos e revisar seu artigo. O único resultado possível desse impasse era a censura de pesquisas críticas de responsabilização.

Dean começou a trabalhar com uma equipe de pesquisadores para escrever um novo artigo que, por fim, revelaria dados reais da emissão de carbono do Google. Para colaborar no trabalho, entrou em contato com Strubell, que havia se tornado professora assistente na Universidade Carnegie Mellon, com uma afiliação parcial à empresa. Depois de inicialmente se animar com a ideia de aumentar a transparência pública sobre o impacto ambiental da IA, ela começou a se perguntar se Dean estaria usando o nome dela para legitimar a crítica ao artigo de Gebru. Um porta-voz do Google disse que Strubell[33] foi convidada porque "correções científicas" muitas vezes são melhores quando o autor dos erros originais participa do processo de correção.

Em uma reunião tensa, Dave Patterson, colaborador de Dean e proeminente pesquisador sênior do Google, enfatizou em termos claros que

seria melhor para a carreira de Strubell participar da pesquisa. Isso lhe daria a chance de corrigir seus erros anteriores e receber crédito por isso. Para ela, as palavras soaram como uma ameaça velada: não participe e o prejuízo será seu. Apesar dos possíveis custos, a alternativa de continuar contribuindo não parecia viável, então Strubell desistiu da colaboração.

A publicação que Patterson fez em um blog[34] sobre o artigo dos pesquisadores do Google em fevereiro de 2022 — intitulada "Good News About the Carbon Footprint of Machine Learning Training" — utilizaria a plataforma da empresa para criticar diretamente o artigo original de Strubell. O estudo de 2019, segundo o post, havia superestimado seriamente as emissões reais do Google para o desenvolvimento do Evolved Transformer em 88 vezes. Essa falha fora causada por dois problemas: o estudo tinha sido realizado "sem acesso imediato ao hardware ou aos data centers do Google" e não havia entendido "as sutilezas" de como a Neural Architecture Search funciona. Como parte da pesquisa que antecedeu a publicação de seus próprios números, os coautores do Google também procuraram Sutskever, seu ex-colega, para obter mais informações sobre o GPT-3. Foi então que a OpenAI[35] e a Microsoft concordaram em divulgar os detalhes técnicos relevantes do modelo pela primeira vez, a fim de calcular seus impactos de consumo de energia e emissão de carbono. Àquela altura, Strubell já estava decepcionada com a indústria e havia deixado a afiliação com o Google. A crítica não acabou prejudicando sua carreira; no entanto, o desgaste emocional da experiência a deixou mais reticente em continuar investigando os impactos ambientais dos grandes modelos de linguagem. Um porta-voz do Google classificou essa situação como "lamentável", acrescentando que "muitos pesquisadores serão necessários para avançar nessa pesquisa — claramente as emissões de carbono são uma preocupação significativa".

Por um breve momento, a reação, os protestos e o dano à reputação do Google pareceram sugerir que um acerto de contas estava em curso. Com o tempo, porém, pesquisadores à procura de emprego e acadêmicos em busca de financiamento já não podiam mais se dar ao luxo de ignorar os vastos recursos financeiros da gigante tecnológica. À medida que a resistência diminuía, a superação do fiasco por parte do Google normalizou um novo processo na empresa para revisões mais abrangentes de pesquisas críticas.

Após o lançamento do ChatGPT, essas normas seriam fortalecidas com a corrida frenética para a comercialização dos sistemas de IA generativa. A OpenAI deixaria principalmente de publicar em conferências de pesquisa. Quase todas as empresas[36] do restante da indústria fechariam o acesso público a detalhes técnicos significativos de seus modelos comercialmente relevantes, que, a partir de então, consideravam patenteados. Em 2023, pesquisadores de Stanford criariam um rastreador de transparência para avaliar empresas de IA quanto à revelação até mesmo de informações básicas sobre seus grandes modelos de aprendizagem profunda, como quantos parâmetros

tinham, com quais dados haviam sido treinados e se havia alguma verificação independente de suas capacidades. Todas as dez empresas[37] avaliadas no primeiro ano, incluindo OpenAI, Google e Anthropic, receberam nota mínima; a pontuação mais alta foi de 54%.

Com essa reversão acentuada nas normas de transparência, a consequência mais alarmante seria a erosão da integridade científica. A base da pesquisa em aprendizagem profunda se apoia em um princípio simples: os dados utilizados para treinar um modelo não são os mesmos utilizados para testá-lo. Sem a capacidade de auditar os dados de treinamento, esse chamado paradigma de divisão entre treino e teste desmorona. Os modelos podem, na realidade, não estar aprimorando sua "inteligência" ao obter pontuações mais altas em diferentes referências, mas sim apenas repetindo as respostas.

CAPÍTULO 8

# O NASCIMENTO DO COMÉRCIO

Mesmo com a crescente controvérsia em relação à abordagem da OpenAI, a determinação da empresa em escalar seus modelos apenas se fortalecia. Para os executivos, o GPT-3 havia definitivamente provado a existência das leis de escala; sendo assim, no início de 2021, estavam prontos para explorar essa fórmula vencedora. O desligamento da equipe da Anthropic também havia diluído a resistência interna mais ferrenha contra a comercialização. Com um novo consenso,[1] a liderança restante elaborou um road map de pesquisa, explicando o foco estreito da empresa e como ele se alimentaria na produção de produtos em um ciclo autorreforçador.

"Nosso objetivo principal para 2021 é desenvolver um sistema alinhado que seja amplamente mais capaz do que qualquer coisa que existia antes", começava o road map. Esse sistema seria, na base, um modelo de linguagem, mas poderia também ser treinado para desenvolver capacidades multimodais. "O objetivo é desafiador, mas conseguimos enxergar um caminho para alcançá-lo em 2021", afirmava o documento.

Esse caminho envolvia três frentes: a primeira era escalar o GPT-3 em mais dez vezes, utilizando um novo supercomputador da Microsoft que chegaria no terceiro trimestre com 18 mil GPUs Nvidia A100, as mais novas e poderosas à época. A segunda era realizar mais pesquisas para aumentar em 25 vezes a eficiência de computação da OpenAI, ou o máximo de poder de processamento que pudesse ser extraído de seus chips disponíveis. A terceira era melhorar a quantidade e a qualidade dos dados de treinamento, em parte aproveitando dados de usuários e direcionando o modelo para as melhores partes da distribuição de dados por meio da aprendizagem por reforço a partir de feedback humano.

Em uma seção intitulada "Como realizar isso", o road map era ainda mais detalhado. Como passo inicial, a OpenAI elevaria diversos modelos de aprendizagem profunda a "uma grande escala", incluindo um modelo de linguagem, um modelo de geração de código, um modelo imagem-para-texto a fim de descrever imagens e um modelo texto-para-imagem a fim de gerar imagens a partir de um prompt de texto. Também seria iniciado um projeto para desenvolver um "agente" digital — um modelo de IA que não apenas

geraria informações semelhantes às humanas, mas poderia receber um objetivo, como enviar um e-mail, e operar de forma autônoma para cumpri-lo. Como próximo passo, a empresa selecionaria um desses modelos para escalar "até o limite permitido pelo nosso cluster de 18 mil A100". Os modelos de linguagem e de código também seriam transformados em produtos e liberados para coletar dados do mundo real dos usuários: "Novidade em 2021: enfatizamos a implementação de modelos como produtos e a aprendizagem a partir das interações do usuário, pois isso pode se tornar uma engrenagem de dados que levará a profundas melhorias de capacidade."

Em outra seção, intitulada "Detalhes", o documento justificava por que essa abordagem fazia sentido tanto científica quanto comercialmente. A experiência anterior da OpenAI demonstrara que aumentar a escala era sua "forma mais confiável de alcançar novas capacidades". Cientificamente, isso significava que a escalabilidade era sua melhor chance de atingir um avanço revolucionário — um tipo de capacidade que antes parecia impossível. Escalar modelos de linguagem e código, em particular, era "tentador pela mera *possibilidade*" de que pudessem alcançar avanços em meta-aprendizagem em nível humano, ou seja, aprender a aprender e raciocinar. Já escalar modelos multimodais poderia potencialmente acelerar ainda mais o ritmo das melhorias. Com os devidos avanços, dizia o documento com notável certeza, "alcançaremos de fato a IAG".

Estrategicamente, como negócio, escalar cada um desses modelos também permitiria "desenvolver capacidades que desejamos utilizar para algum fim". Modelos de linguagem e código melhores poderiam tornar a própria OpenAI mais produtiva e acelerar seu progresso. Juntamente com modelos aprimorados de texto-para-imagem, eles também "levariam a produtos incríveis".

Em paralelo, a OpenAI precisava investir em eficiência computacional. Embora a escalabilidade tivesse funcionado maravilhosamente para manter a empresa na primeira posição, a estratégia começava a demonstrar sinais de desaceleração. "Nos últimos dois anos, fizemos progressos extraordinários por meio da escalabilidade apenas porque havia um grande *excedente de hardware*", dizia o documento, referindo-se ao fato de que pesquisadores de inteligência artificial não haviam utilizado antes o número máximo de chips de computador disponíveis para treinar modelos de IA. Como tal, "conseguimos superar muito o restante do mundo de ML ao usarmos todo o poder computacional disponível para treinar modelos de tamanho e capacidade então sem precedentes".

Nessa época, a OpenAI estava "aproximando-se do limite" da quantidade de poder computacional que poderia adquirir. A empresa também enfrentava nova concorrência de "outros laboratórios" que haviam adotado a mesma estratégia de escalabilidade, fato reconhecido no documento, mas sem mencionar explicitamente a Anthropic. "Nossa capacidade de expansão é tal que, nos próximos dois anos, seremos capazes de treinar um modelo que use cem vezes o poder computacional do GPT-3." Embora a escalabilidade por si

só produzisse progressos "formidáveis", não se compararia ao salto do GPT-2 para o GPT-3, que fora impulsionado por um aumento de quinhentas vezes no poder computacional. "Todo progresso adicional deve vir de métodos melhores", concluía o road map.

O documento listava diversas áreas de exploração para identificar esses métodos. Alguns eram chamados de métodos "2x" e "10x" — aqueles que poderiam potencialmente alcançar ganhos de duas ou dez vezes em eficiência computacional. As sugestões incluíam destilação, a engenharia reversa de modelos menores a partir de modelos maiores; filtragem de dados, identificando aqueles que proporcionariam os maiores saltos de desempenho; e esparsidade, desenvolvendo modelos de IA considerados esparsos, ou mais leves. Este último se referia a uma característica das redes neurais: em um modelo tradicional de aprendizagem profunda, uma rede neural é "densamente" conectada, com cada nó de uma camada ligado a todos os nós de outra camada. Modelos esparsos são treinados com apenas um pequeno subconjunto dos nós conectados e com o objetivo de reduzir significativamente os custos computacionais em troca de modelos ligeiramente menos precisos, mas ainda bons o bastante para a maioria dos propósitos.

Além dos métodos de duas ou dez vezes, a equipe de Pesquisa precisava buscar outros meios que pudessem "acentuar a inclinação da lei da escalabilidade" — ou seja, produzir saltos maiores no desempenho do modelo sem aumentar seus dados, parâmetros ou poder computacional. A melhor hipótese da OpenAI até então para os métodos mais promissores, segundo o documento, era mediante o raciocínio e a aprendizagem ativa — uma técnica em que um modelo de IA identifica repetidamente quais partes de um conjunto de dados devem ser priorizadas para anotação por trabalhadores humanos. (Pouco depois, um grupo de pesquisadores da empresa descobriria um pequeno erro nas leis de escalabilidade originais, o qual indicava que a OpenAI precisaria treinar seus modelos um pouco mais do que acreditava inicialmente para obter melhor desempenho. Com uma pitada de soberba, o grupo divulgou a descoberta como uma vantagem competitiva única: "Isto é algo que temos agora e que a Anthropic não tem." Um ano depois, o Google[2] publicaria um artigo revelando o mesmo resultado.)

Por fim, o road map acrescentava que a OpenAI precisava começar a buscar "o sistema inovador do futuro" — um avanço tão significativo quanto o GPT-1, que pudesse fornecer à empresa um novo caminho de desenvolvimento a ser explorado. Talvez isso surgisse do trabalho com escalabilidade e eficiência computacional, mas a empresa continuaria a realizar pesquisas exploratórias na vanguarda da área, incluindo o desenvolvimento de algoritmos para resolver problemas de matemática e experimentos com sistemas multiagentes, e a exploração de outras ideias novas. Como parte desse trabalho, os pesquisadores continuariam a "estudar a ciência da aprendizagem profunda para entender melhor de que maneira nossas ferramentas realmente funcionam". Em outras palavras, a OpenAI precisava compreender com mais

exatidão o que de fato estava construindo.

À medida que a equipe de Pesquisa avançava seguindo seu road map, a subdivisão da área de Produtos começou a estruturar suas forças, contratando um líder de Go-to-Market, uma equipe de vendas e mais engenheiros. Ao atender uma base crescente de desenvolvedores, a API do GPT-3 se tornou o campo de testes para resolver os problemas de produtização e monetização, incluindo a adaptação da infraestrutura de back-end da empresa para suportar um serviço com usuários e a criação de uma estratégia de precificação.

A disponibilização do serviço aos usuários também levantou questões sobre quais tipos de comportamentos a OpenAI permitiria ou não em seus produtos, assim como novas responsabilidades para garantir o cumprimento dessas regras. Na época, a empresa ainda não tinha uma equipe oficial de Trust and Safety, uma área consolidada na indústria de tecnologia — não confundir com as preocupações existenciais do clã de Segurança da OpenAI — para lidar com tais questões e antecipar e prevenir uma ampla gama de abusos na internet, como lavagem de dinheiro, cyberbullying e desinformação. Em vez disso, a OpenAI contratou um pequeno grupo de funcionários e prestadores de serviços para revisar os pedidos de acesso à API realizados pelos desenvolvedores e aprová-los ou rejeitá-los com base em regras improvisadas criadas ao longo do caminho. Muitos dos limites definidos eram arbitrários. A empresa decidiu aceitar bots de acompanhantes, mas não bots sexuais; permitir apps de geração de textos para redes sociais, mas não aqueles que postassem de modo direto ou se passassem por figuras públicas. "Tudo era praticamente baseado na intuição", disse uma pessoa envolvida na definição das diretrizes. "Havia muita tentativa e erro enquanto revisávamos."

Outra equipe, liderada por Ari Herbert-Voss, cientista de pesquisa com formação em segurança que havia se juntado à empresa em 2019, buscava identificar e corrigir os comportamentos indesejáveis do GPT-3 e suas respostas propensas a erros, tentando explorar e "quebrar" o modelo de diversas formas e, em seguida, projetando mecanismos para torná-lo mais resistente a falhas e ao uso indevido. Um desses mecanismos incluía o desenvolvimento de uma versão inicial do filtro de moderação de conteúdo, para o qual a OpenAI contrataria mais adiante trabalhadores no Quênia. Os pesquisadores treinavam o filtro com todos os exemplos que conseguiam encontrar ou que podiam gerar e escrever a partir dos próprios modelos de IA. No entanto, quando lançado, o filtro funcionava indevidamente, bloqueando amplos trechos de conteúdo benigno, como referências básicas a pessoas negras ou trans. Desenvolvedores e outros usuários da API se queixaram disso. Diversos membros da equipe da API já tinham ressalvas quanto a aplicar qualquer filtragem, preocupados com o impacto na experiência do cliente. Sendo assim, tornaram o filtro opcional.

A OpenAI chamou esse processo[3] de testar e refinar o modelo de *red teaming*, um termo apropriado da indústria de cibersegurança que se refere a

um processo sistemático e rigoroso de verificação da segurança de uma organização e de sua capacidade de responder a um ataque. A versão de *red teaming* da OpenAI, segundo Heidy Khlaaf, engenheira de segurança, especialista em cibersegurança e pesquisadora de IA, não era e ainda não é isso. É parcial e improvisada, além de não estabelecer nenhuma garantia sobre a segurança do modelo. Khlaaf, que trabalhou com a OpenAI[4] bem no início, quando a empresa tentava estabelecer protocolos de teste de estresse, ficou alarmada ao ver como a indústria de IA se apropriou de uma terminologia já consolidada em sua área para criar uma aparência falsa de rigor. "Em engenharia de software, testamos muito mais uma simples calculadora", ela observou. (Sim, uma calculadora.) "Não é por acaso que eles usam uma terminologia que carrega muita credibilidade."

Um dos primeiros clientes que gerou uma discussão interna significativa foi a Luka, uma empresa de São Francisco que desenvolvia um aplicativo de amigos virtuais criados a partir de IA chamado Replika. A empresa havia se associado[5] à OpenAI para o lançamento da API do GPT-3 com o objetivo de melhorar a fluidez das conversas do produto. Apesar do branding de *companion bot* do Replika, a OpenAI logo descobriu que os usuários do aplicativo com frequência participavam de conversas sexualmente explícitas. Funcionários debateram se isso se enquadrava na linha de aceitabilidade. No fim, a empresa decidiu proibir o Replika de utilizar seu modelo. Além das preocupações com conteúdo sexual, o aplicativo alimentado pelo GPT-3 às vezes gerava respostas emocionalmente manipuladoras, fazendo com que os usuários acreditassem que seu Replika, assim como um humano, poderia se sentir magoado se eles não interagissem regularmente. A equipe da OpenAI também começou a se sentir desconfortável por poder ler essas conversas.

Em outro momento, Brockman deu acesso à API à Latitude, uma startup de Utah onde seu irmão trabalhava e que criava mundos virtuais com IA. A empresa já utilizava[6] um modelo anterior da OpenAI para alimentar um jogo de escolhas inspirado em Dungeons and Dragons, no qual os usuários podiam digitar qualquer ação que quisessem executar em uma caixa de diálogo. Com a nova API, a Latitude atualizou seu jogo para rodar no GPT-3. Poucos meses depois, alguns usuários começaram a gerar cenários textuais envolvendo abuso sexual infantil. Após descobrir o problema por meio do sistema de monitoramento relativamente novo da OpenAI, Ari Herbert-Voss levou a questão ao restante da equipe.

"Encontrei umas coisas ontem à noite. Há muito conteúdo sexual sendo gerado", disse Herbert-Voss em uma reunião.

No começo, as pessoas riram. "Certo, mas é assim que a internet funciona, não é?"

"Não, isso é material de abuso sexual infantil", respondeu ele.

O pânico se alastrou. "Droga, como vamos evitar isso?"

O incidente gerou uma longa negociação entre a OpenAI e a Latitude sobre como lidar com a situação. Brockman se preocupava em tomar medidas

punitivas que poderiam afetar severamente a empresa de seu irmão. No fim, a Latitude implementou às pressas um filtro para bloquear o conteúdo de abuso sexual infantil em formato de texto. A OpenAI emitiu um comunicado público para se distanciar da falta de moderação de conteúdo e, de forma sutil, culpabilizar a Latitude. No entanto, para alguns funcionários da OpenAI, a culpa estava claramente na própria tecnologia e na ausência de processos, e o incidente pesou bastante. A Latitude já havia banido alguns usuários por gerarem conteúdo sexual envolvendo crianças com o modelo anterior da OpenAI; era previsível que isso aconteceria novamente, em escala, com o GPT-3. "Foi triste para mim termos lançado essa API com nossa missão de beneficiar a humanidade, e todos tiveram impressões muito positivas sobre como nossos usuários estavam economizando tempo em atendimento ao cliente ou qualquer coisa do gênero", disse um ex-funcionário da OpenAI, "mas na realidade, muito do nosso tráfego estava indo para conteúdo sexual infantil no AI Dungeon e para um produto assustador de namoradas virtuais."

De volta à divisão de Pesquisa, a equipe de geração de código era a que avançava mais rapidamente.

"O Divórcio" resultara na unificação dos esforços duplicados de geração de código, e Wojciech Zaremba havia se tornado o principal responsável. Cientista da computação polonês que cresceu vencendo competições de matemática, programação, química e física, Zaremba era conhecido por sua incrível aptidão técnica e atenção à formação de equipes, além de sua paixão pelo poder curativo da amizade, da natureza, do sexo e das drogas. Às vezes, contava com orgulho para pessoas no escritório sobre planos para retiros que durariam semanas. "Vamos caminhar por quase 13 quilômetros. Em seguida, vamos transar e depois caminhar por mais 13 quilômetros", disse, gabando-se para outro líder da OpenAI, enquanto outros ao redor ouviam constrangidos.

Ainda no auge da pandemia, Zaremba havia pedido que sua equipe de aproximadamente dez pesquisadores comparecesse ao escritório, mesmo enquanto outras trabalhavam remotamente, por acreditar que o trabalho presencial se fazia necessário para resolver o desafio do desenvolvimento do modelo. Depois de ver as capacidades de geração de código do GPT-3, Murati sugeriu a Kevin Scott, CTO da Microsoft, a ideia de transformar essas habilidades em um produto de assistência de programação com IA. Em 2018, a Microsoft havia adquirido o GitHub, a plataforma mais popular para desenvolvedores armazenarem e compartilharem seus códigos e da qual a OpenAI já vinha extraindo dados por conta própria. Os executivos da Microsoft orientaram[7] a plataforma a entregar à OpenAI todo o código de seus repositórios públicos a fim de evitar trabalho. À medida que a equipe de geração de código obtinha resultados cada vez melhores, Altman comparecia regularmente às reuniões, incentivando os pesquisadores a prosseguirem e entregarem o que tivessem de melhor à Microsoft. Na primavera, após empolgar os executivos da gigante tecnológica com diversas demonstrações,

ficou claro que o modelo seria o segundo projeto comercial da OpenAI, depois do GPT-3.

Alguns membros da equipe de Scott tinham ressalvas sobre o desenvolvimento do modelo. Embora dar à OpenAI acesso gratuito ao código dos repositórios públicos do GitHub não fosse ilegal, ainda parecia uma violação da confiança da comunidade de usuários. Grande parte daqueles códigos havia sido compartilhada com o objetivo de fomentar o desenvolvimento de software open-source, a fim de ajudar desenvolvedores independentes e pequenas startups a terem chance de competir, e não de ajudar grandes empresas a fortalecerem seu monopólio. Em um memorando, eles apresentaram suas críticas principais ao projeto do GitHub, sugerindo que Scott reconsiderasse a ideia de sugar os dados de desenvolvedores publicados sob uma licença Creative Commons sem consentimento ou compensação, lembrou um ex-funcionário. A Microsoft, segundo o memorando, deveria considerar cancelar o produto ou, no mínimo, destinar uma porcentagem dos lucros e devolvê-la à comunidade open-source. Embora Scott estivesse receptivo à ideia, criar a ferramenta e ser o primeiro no mercado era seu foco central, disse o ex-funcionário. No fim, a Microsoft doou certa quantia a um programa já existente de apoio a desenvolvedores open-source chamado GitHub Sponsors e deixou a visão do produto inalterada.

Dentro da OpenAI, os funcionários justificavam o projeto com diferentes argumentos. Alguns concordavam com Altman, defendendo que trabalhar em um produto para agradar a Microsoft e, assim, continuar garantindo dinheiro e recursos computacionais parecia primordial para cumprir a missão da empresa. Para outros, um modelo de geração de código parecia altamente valioso economicamente, alinhando-se bem com a definição da OpenAI de que a IAG consistia em "sistemas altamente autônomos que superam humanos na maior parte do atividades com maior valor econômico". Nesse sentido, Altman também via a geração de código como uma forma de acelerar a própria atividade com maior valor econômico da OpenAI, uma convicção que mais tarde alimentaria o início de um projeto chamado AI Scientist, voltado para o avanço dos modelos da OpenAI para a realização de pesquisas em IA de forma autônoma.

Para muitos pesquisadores, havia também um terceiro argumento: a dedicação era um passo importante para desenvolver o próximo modelo GPT, o qual esperavam que fosse capaz de realizar certo grau de raciocínio, ingrediente-chave ainda ausente. No campo mais amplo, enquanto debates ferviam entre as correntes de Hinton e Marcus sobre se a aprendizagem profunda por si só seria capaz de produzir um modelo com tal habilidade, os pesquisadores da OpenAI levantavam a hipótese de que, caso fosse possível, treinar um modelo com código provavelmente ajudaria. Dados de programação eram uma das fontes mais óbvias e volumosas de padrões estruturados de lógica. O argumento voltava sempre à mesma origem: se a geração de código ajudasse a fazer os modelos de IA avançarem rumo à IAG,

que maneira melhor haveria de cumprir a missão da OpenAI?

Apesar das bilhões de linhas de código disponíveis no GitHub, o volume de dados ainda era pequeno em comparação ao que havia sido utilizado para treinar o GPT-3. A equipe acreditava que o modelo de geração de código precisaria ser treinado *tanto* com o GitHub *quanto* com o conjunto de dados do GPT-3 para obter os melhores resultados. Também encontraram novas fontes de dados, incluindo raspagens do Stack Overflow, um fórum on-line semelhante ao Quora onde desenvolvedores publicam perguntas para a comunidade, manuais de instruções de programação e livros didáticos de programação em diferentes linguagens. A dúvida era como combinar esses dados da melhor maneira: seria preferível ajustar o GPT-3 existente com o GitHub e outros materiais ou treinar um modelo novo do zero, com tudo novo mesclado ao antigo? A equipe optou pelo fine-tuning para economizar dinheiro; os experimentos que estavam realizando já custavam algo em torno de 100 mil dólares cada um, com base nos preços da Microsoft para seus serviços em nuvem. Treinar um novo modelo poderia custar dezenas de milhões. Mais tarde, ao desenvolver o que a OpenAI chamaria de GPT-3.5, a abordagem foi alterada, misturando os dados desde o início. Pelas métricas internas da OpenAI, os resultados realmente sugeriam que a adição de dados de programação havia melhorado a capacidade do modelo de realizar tarefas baseadas em lógica, não apenas em código, mas também em inglês — um fenômeno conhecido como aprendizagem por transferência.

No verão de 2021, a OpenAI entregou o Codex, uma versão inicial e bruta de seu modelo de geração de código, ao GitHub e à Microsoft. O modelo era grande e lento demais, tornando-se caro para operar em escala e gerando uma experiência negativa para o usuário. Tensões surgiam à medida que as três organizações lidavam com os desafios de sua primeira colaboração. Havia confusão sobre quem seria o responsável — a OpenAI ou o GitHub — por otimizar o modelo para que se tornasse um produto utilizável. Além disso, faltava clareza entre os funcionários da OpenAI sobre quanto de propriedade intelectual deveriam compartilhar com seus colegas do GitHub, enquanto os funcionários do GitHub tinham dificuldade de saber até que ponto confiar na OpenAI. Os conflitos se acumulavam conforme as empresas divergiam a respeito de como e quando lançar o produto e quem receberia o crédito.

Por fim, Murati acabou negociando um acordo — uma habilidade que lhe renderia crescente respeito entre aqueles que trabalharam com ela em diferentes empresas. A Microsoft teria seu momento[8] ao lançar, em julho de 2021, o GitHub Copilot, seu produto voltado ao consumidor. A OpenAI então lançaria[9] a própria versão do Codex diretamente na API da empresa em agosto.

O acordo daria[10] à Microsoft uma nova base de usuários e um impulso financeiro modesto: em dois anos, o GitHub Copilot chegaria a um milhão de assinantes pagantes, gerando mais de 100 milhões de dólares em receita anual recorrente. No entanto, para a OpenAI, o acordo aprofundou uma sensação

crescente de que seria melhor trabalhar nos próprios produtos voltados ao consumidor. Os pesquisadores da empresa haviam se dedicado intensamente ao modelo e estavam cedendo todo o reconhecimento de marca ao GitHub e à Microsoft, e ver ambas as empresas aproveitarem publicamente o crédito pelo trabalho da OpenAI foi algo difícil de digerir. A Microsoft também era um parceiro desafiador; muitos sentiam que tinha burocracia demais e exigia apoio excessivo para aproveitar ao máximo os modelos da OpenAI. Ao depender da gigante tecnológica para entregar suas tecnologias ao público, a OpenAI também estava perdendo visibilidade, dados de seus usuários e, mais importante, controle sobre sua visão.

À medida que a OpenAI concentrava suas apostas, Altman aplicava a mesma estratégia a seus outros projetos e investimentos. Ao longo dos anos, conforme se voltava mais para inovação em hard tech, ele desenvolveu a convicção de apostar alto e por longos períodos nos projetos mais importantes.

"Se você pudesse agitar[11] uma varinha e alterar qualquer coisa no ecossistema de tecnologia, startups e empreendedorismo, o que alteraria?", seu irmão Jack lhe perguntou em um evento, quando Altman deixou a YC.

"Seria fazer com que todo mundo no ecossistema adotasse um horizonte temporal muito mais longo", Altman respondeu. "Esse mundo onde as pessoas começam uma empresa e planejam tocá-la por quatro ou cinco anos, ou entram em uma organização e já planejam sair em um ou dois — não é assim que coisas importantes são feitas."

Dessa forma, 2021 pareceu marcar o momento em que Altman realizou uma grande mudança na estratégia de investimento pessoal, passando de um grande número de pequenas apostas ao investimento em um pequeno número de apostas realmente grandes. Naquele ano, a Tools for Humanity, uma startup que ele havia cofundado em 2019 e que até então permanecera discreta, recebeu um influxo de financiamento e cobertura da mídia, além de uma aceleração coordenada em suas operações, conforme Altman direcionava mais atenção externa para a empresa. O empreendimento era uma tentativa[12] de desenvolver um mecanismo funcional para a renda básica universal, uma ideia popular no Vale do Silício de oferecer a todos uma distribuição mínima regular de renda. Altman falava frequente e apaixonadamente sobre a renda básica universal como possível antídoto para um futuro em que a IA poderia causar impactos econômicos intensos. Na YC, ele havia iniciado[13] o maior piloto do país para estudar o conceito, criando no processo a OpenResearch, uma organização sem fins lucrativos, que administrou o programa. Ao longo de três anos, a OpenResearch deu um subsídio mensal de mil dólares a um grupo de mil pessoas de baixa renda, escolhidas aleatoriamente dentre 3 mil, enquanto o restante recebia cinquenta dólares por mês como grupo controle. Em julho de 2024, a organização[14] divulgaria os resultados, mostrando que a renda garantida ajudou as pessoas a atenderem às necessidades básicas, ajudar outras e ter maior margem econômica.

O principal produto da Tools for Humanity,[15] o Worldcoin, era uma criptomoeda que se autodeclarava "pertencer ao coletivo" e que permitiria que todos, porventura, recebessem uma parcela de seu valor. Como parte do esquema, a empresa estava desenvolvendo um orbe cromado e de aparência chamativa — aproximadamente do tamanho de uma bola de boliche e, em parte, um reflexo do gosto estético de Altman — para escanear a íris e verificar a identidade das pessoas antes de entregar a parte de cada uma. Os fundadores argumentavam que o escaneamento de íris seria uma necessidade assim que a IA dificultasse ainda mais a distinção do que era mídia falsa e o que era real. Mais tarde, uma extensa investigação[16] de Eileen Guo e Adi Renaldi na *MIT Technology Review* descobriria que essas tentativas de escaneamento de íris estavam envolvidas em violações de privacidade de dados, práticas de marketing enganosas e possíveis infrações legais. Em julho de 2023, a Worldcoin[17] seria lançada oficialmente em meio a grande controvérsia, quando milhares de pessoas começaram a formar filas, principalmente em países do Sul Global, para entregar seus dados biométricos sem entender ao certo para que estavam fazendo aquilo, além da vaga promessa de dinheiro gratuito.

Também em 2021, Altman fez seus dois maiores investimentos: 180 milhões de dólares na Retro Biosciences, uma empresa de biotecnologia que trabalha para estender a vida humana por meio do rejuvenescimento celular, e 375 milhões na Helion Energy, que comercializa a fusão nuclear. "Eu basicamente peguei[18] todo o meu patrimônio líquido disponível e coloquei nessas duas empresas", Altman disse a Antonio Regalado, da *MIT Technology Review*. Ele descreveu ambas as tecnologias em uma linguagem que refletia o roteiro de pesquisa da OpenAI — pareciam impossíveis no momento, mas, se escaladas agressivamente, logo poderiam se tornar realidade.

A aposta na Retro Biosciences refletia a fixação de Altman com a longevidade. Era um seguidor ávido da pesquisa sobre "sangue jovem" — uma linha de investigação científica que estudava como reverter o envelhecimento com transfusões de sangue mais saudável e mais jovem. Perceptivelmente, era uma área na qual Thiel também tinha interesse, gerando uma infinidade de artigos e memes sobre seu desejo de injetar em si mesmos sangue de adolescentes. Enquanto estava na YC, Altman também havia se inscrito mediante um depósito de 10 mil dólares para estar na lista de espera de uma startup controversa chamada Nectome, que havia participado de uma das turmas da incubadora de Altman. Parecendo ter sido extraída diretamente da ficção científica, a Nectome estava oferecendo um serviço que congelaria criogenicamente os cérebros dos clientes para um dia — potencialmente centenas de anos no futuro — serem carregados em um computador depois que os cientistas tivessem desenvolvido a tecnologia necessária. O problema era que a Nectome precisava que o cérebro da pessoa estivesse fresco para que a preservação funcionasse. Para Antonio Regalado, o cofundador[19] Robert McIntyre chamou seu produto de "100% fatal".

A Helion refletia a obsessão de Altman em encontrar maneiras de gerar energia limpa, barata e abundante. Com frequência, ele observava que o custo e a disponibilidade de energia estavam extremamente correlacionados com a qualidade de vida e o crescimento econômico. Contudo, sem alternativas livres de carbono, o aumento do consumo de energia "destruiria o planeta",[20] segundo ele. Em 2023, Altman descreveria a Helion como "mais do que um investimento"[21] e "a outra coisa além da OpenAI na qual eu gasto muito tempo". A Microsoft subsequentemente assinaria um acordo para comprar energia da primeira usina da Helion, depois de ter feito seu terceiro investimento na OpenAI, no valor de 10 bilhões de dólares. Para o espanto[22] e ceticismo de especialistas em energia, a Helion se comprometeria a ter sua usina pronta até 2028.

Foi também em 2021 quando Altman levou sua propensão a investir para a OpenAI. Em maio, lançou[23] o OpenAI Startup Fund, um fundo de investimento de 100 milhões de dólares para apoiar empresas em estágio inicial com, conforme ele descreveu, "grandes ideias sobre como utilizar a IA para transformar o mundo". Mais uma vez, a Microsoft se tornou investidora no fundo. Para alguns observadores, sua criação foi uma decisão estranha. A OpenAI mal estava gerando receita e já era altamente dependente de capital intensivo. Por que levantar ainda mais dinheiro para ser usado por empresas separadas? Outros sentiram que essa era a forma de Altman replicar os poderosos efeitos de network da YC na OpenAI. Ainda houve aqueles que viram simplesmente como um costume de Altman. "Esta é a forma como Sam se movimenta pelo mundo", disse uma pessoa que trabalhou com ele. "Fechando acordos."

Altman gostava de falar que não tinha participação acionária na OpenAI para evitar corromper a busca pela IAG segura com os próprios desejos de lucro. Recebia apenas um salário anual de 65 mil dólares e acumulava sua riqueza por meio de outros empreendimentos, o que soava bem — e confirmava sua retórica original sobre o porquê de a OpenAI ter começado como uma organização sem fins lucrativos. Também era uma afirmação que, igual ao status de organização sem fins lucrativos da empresa, já não refletia toda a verdade em 2021. Altman tinha uma participação significativa na YC, e esta, por meio de seu investimento de 10 milhões de dólares na OpenAI, poderia receber até um bilhão em retorno. À medida que a OpenAI continuava a se comercializar, muitas startups da incubadora e muitos de seus outros investimentos também se tornariam clientes ou parceiros comerciais da empresa. Em julho de 2024, o *Wall Street Journal* calcularia o patrimônio líquido de Altman,[24] considerando todos os seus investimentos, em pelo menos 2,8 bilhões de dólares. Com o OpenAI Startup Fund, Altman acrescentou mais uma complicação à narrativa altruísta — que acabaria desempenhando o próprio papel em sua breve derrocada.

CAPÍTULO 9

# O CAPITALISMO DO DESASTRE

À medida que a OpenAI avançava a todo vapor, conduzida pelas convicções de Altman, os limites das consequências tão abrangentes da visão da empresa se expandiam. Com a alimentação de seus modelos com conjuntos de dados cada vez maiores e mais poluídos, havia sido criada a "mudança de paradigma" que Ryan Kolln, da Appen, me descreveria — o foco de filtragem das entradas de dados foi alterado para o controle das saídas do modelo. A linguagem da abstração mais uma vez ocultava uma realidade sombria: o que essa mudança realmente significava para as pessoas que então precisavam arcar com o peso do novo foco.

Em 2021, paralelamente ao esforço de desenvolver as próximas gerações de seus modelos, a OpenAI iniciou um projeto para criar uma versão muito mais aprimorada de seu filtro automatizado de moderação de conteúdo, utilizado para limpá-los.[1] Enquanto o GPT-3 havia sido colocado na API sem qualquer filtragem, levando ao escândalo de pornografia infantil baseada em texto envolvendo a Latitude, a empresa queria ser mais cuidadosa com os modelos que passaria a chamar de GPT-3.5 e, posteriormente, GPT-4. À medida que a OpenAI se preparava para implantar suas tecnologias de forma mais ampla, ter um produto completamente sem filtro poderia se revelar problemático no longo prazo, sob os aspectos jurídicos, de relações públicas e de usabilidade. Na época, os planos para o que se tornaria o ChatGPT ainda não haviam sido concebidos, mas o chatbot também se beneficiaria futuramente do mesmo filtro. Este atuaria como uma camada protetora em torno de cada modelo, com o intuito de sinalizar e remover de suas saídas conteúdos ofensivos antes que chegassem ao usuário.

Para construir o filtro automatizado, primeiro a OpenAI precisava de trabalhadores humanos que pudessem revisar e catalogar meticulosamente centenas de milhares de exemplos do tipo de conteúdo exato — sexo, violência e abuso — que a empresa queria impedir que seus modelos gerassem.[2] Depois de seis meses de busca,[3] encontrou um fornecedor que parecia adequado para assumir o projeto: uma empresa de terceirização que vinha realizando moderação de conteúdo para a Meta desde 2019 e que, por

coincidência, compartilhava o apelido de Altman, a Sama. A OpenAI enviou-lhes um e-mail[4] perguntando se a empresa aceitava projetos envolvendo conteúdo sensível ou explícito e qual era sua abordagem típica para lidar com isso. Após receber respostas detalhadas,[5] a OpenAI assinou quatro contratos[6] com a Sama no valor de 230 mil dólares, colocando o projeto nas mãos de dezenas de trabalhadores no Quênia.

Não é coincidência que o Quênia tenha se tornado o local de um dos tipos mais exploratórios de trabalho envolvidos na criação do ChatGPT. Havia anos que o país estava entre os principais destinos para os trabalhos mais pesados e terceirizados pelo Vale do Silício. Com muitos outros países relegados a esse papel pela indústria de tecnologia, o Quênia compartilha um denominador comum: é pobre, está localizado no Sul Global e conta com um governo ávido por investimentos estrangeiros vindos de países mais ricos. Todos esses fatores fazem parte do legado colonial do país, que o deixou sem instituições bem desenvolvidas para proteger seus cidadãos da exploração, que se veem frequentemente em meio a crises econômicas, circunstâncias que criam um terreno fértil para que empresas estrangeiras encontrem uma massa de trabalhadores empobrecidos dispostos a realizar tarefas por peça sob quase quaisquer condições.

É possível ver os sinais desse legado nos inúmeros rostos de Nairóbi, a capital do Quênia. A cidade sofre com a grave desigualdade.[7] O distrito comercial central abrange torres reluzentes, hotéis internacionais cinco estrelas e restaurantes de alto padrão. O bairro diplomático tem edifícios magnificentes e muros altos de segurança. As áreas residenciais de expatriados oferecem mansões extravagantes com jardins privados deslumbrantes. E então vemos as periferias: Utawala, Dagoretti Sul, Embakasi. Vá a qualquer um desses bairros, e os arranha-céus de aço são substituídos por estruturas baixas de blocos de concreto; os prédios começam a se espalhar como ervas daninhas, em padrões irregulares, sem planejamento coordenado; as ruas passam de asfaltadas a não asfaltadas, de quatro faixas a estreitas, destinadas principalmente a motocicletas e pedestres; e, mais adentro, o concreto cede lugar a chapas de metal ondulado, e casas e comércios improvisados se amontoam cada vez mais.

Nessas condições, o governo do Quênia abraçou de bom grado a chegada do Vale do Silício em busca de trabalhadores de baixa remuneração.[8] O país conta com uma indústria local limitada. Muitas das melhores marcas são europeias e americanas. Algumas das maiores infraestruturas, antes construídas pelos britânicos, agora são construídas pelos chineses. Os carros são, em grande parte, herdados do Japão, onde os motoristas também se sentam do lado direito do veículo. Aos olhos do governo, as gigantes tecnológicas poderiam ajudar o país a criar os empregos dos quais tanto precisava. O desemprego gera crime, e pequenos furtos são comuns. Pessoas que se sentem desamparadas tornam-se desconfiadas em relação às instituições. Durante a guerra entre Rússia e Ucrânia, ao passo que os preços

dos grãos subiam no Quênia, rumores se espalhavam de que o presidente estaria intencionalmente pressionando famílias que já passavam fome. Muitos repetiram um refrão familiar nos EUA: a eleição foi fraudada.

E, assim, o Quênia se tornou um ponto crucial para a mão de obra de retaguarda da internet. Diversas empresas como a Sama — intermediárias na cadeia de fornecimento de trabalho de dados — estabeleceram operações em Nairóbi, formando grupos de trabalhadores para atender empresas de tecnologia estrangeiras, principalmente na Bay Area.

Para a OpenAI, a Sama parecia atender a todos os requisitos. Originalmente chamada Samasource, era uma empresa social sediada em São Francisco, fundada em 2008 com a missão de oferecer trabalho significativo e digno a pessoas em países empobrecidos, ajudando-as a sair da pobreza. Sob a liderança de Leila Janah, sua fundadora, estabeleceu operações na Índia e no Quênia, desenvolvendo reputação como uma empresa de terceirização ética. Em 2018, tornou-se uma empresa com fins lucrativos, encurtando seu nome para expandir as operações. Em 2020, recebeu a certificação B Corp. Nas respostas às perguntas da OpenAI em 2021, a organização detalhou sua experiência em moderação de conteúdo e enfatizou seus protocolos para manter os projetos em sigilo e proteger seus dados. Mencionou também que oferecia recursos de saúde mental aos trabalhadores para ajudá-los a lidar com os conteúdos psicologicamente perturbadores.

Entretanto, nos bastidores, o caos tomava conta da Sama. Em janeiro de 2020, Leila Janah faleceu de um câncer raro, aos 37 anos; somada à pandemia que veio logo em seguida, sua morte marcou, segundo trabalhadores, o início de uma fase de má administração da empresa — o que foi negado por um porta-voz da empresa. Os problemas viriam à tona somente no início de 2022, quando Billy Perrigo, repórter da revista *Time*, publicou uma extensa investigação[9] na qual revelou que a Sama havia assumido um projeto para a Meta, moderando o conteúdo do Facebook para toda a África Subsaariana, expondo repetidamente os trabalhadores a vídeos violentos e explícitos — incluindo suicídios e decapitações —, traumatizando-os de modo profundo e causando sofrimento. A Sama se defendeu,[10] declarando ter aceitado o projeto após uma cuidadosa deliberação da equipe da África Oriental, que queria garantir que "o conteúdo voltado para africanos fosse revisado por africanos". Cerca de duzentos trabalhadores abririam múltiplos processos contra a Sama e a Meta, alegando condições de trabalho traumáticas e demissões ilegais por tentarem se organizar em busca de salários mais altos e melhores condições.[11] O porta-voz da Sama negou as acusações.

Foi nesse contexto que a OpenAI iniciou a primeira fase de seu projeto, no fim de 2021. Sob os codinomes PBJ1, PBJ2, PBJ3 e PBJ4,[12] equipes de trabalhadores da Sama foram lançadas em tarefas de moderação de conteúdo ainda mais traumáticas, recebendo, em média, entre 1,46 e 3,74 dólares por hora. Os trabalhadores não faziam ideia para quem ou por que estavam realizando o projeto, sendo mantidos no escuro sob cláusulas de

confidencialidade, uma prática comum na indústria de anotação de dados.[13] O que sabiam era apenas o que tinham diante de si: centenas de milhares de descrições textuais grotescas que precisavam ser lidas e classificadas de acordo com a gravidade.[14] Seria violência ou violência extrema? Assédio ou discurso de ódio? Abuso sexual infantil ou zoofilia?

Aos poucos, o trabalho destruiu a saúde mental de muitos dos trabalhadores, e os impactos se espalharam além de cada indivíduo, atingindo também as pessoas que dependiam deles em suas comunidades. Somente depois do lançamento do ChatGPT é que eles começaram a entender o que exatamente haviam pagado com sua sanidade. Em maio de 2023, visitei quatro trabalhadores em Nairóbi que concordaram em compartilhar suas experiências comigo publicamente, para uma reportagem de capa do *Wall Street Journal*. Um deles, Mophat Okinyi, fazia parte da equipe responsável por conteúdo sexual. O projeto que tanto abalou sua mente e suas relações interpessoais acabou se revelando um trabalho a serviço de uma tecnologia que, ironicamente, contribuiria para a erosão das oportunidades econômicas de seu próprio irmão.[15]

Em uma época na qual ainda era permitido realizar pesquisas corporativas livremente, Mary L. Gray, antropóloga da Microsoft, e Siddharth Suri, cientista social computacional, estavam entre os primeiros a revelarem ao mundo a situação de trabalhadores como aqueles do Quênia — pessoas cuja subsistência dependia de uma engrenagem essencial da cadeia produtiva da IA.

Em 2019, eles publicaram o livro *Ghost Work*,[16] resultado de cinco anos de extensa pesquisa de campo, expondo uma teia oculta de trabalho fragmentado e exploração digital que sustentava o Vale do Silício. As gigantes e os unicórnios da tecnologia construíam suas avaliações bilionárias não apenas com engenheiros que recebiam salários de seis dígitos em escritórios descolados; eram igualmente fundamentais os trabalhadores, muitas vezes do Sul Global, aos quais era paga uma ninharia para anotar meticulosamente enormes volumes de dados.

Pensemos no caso dos carros autônomos. Um veículo desses precisa dirigir na faixa correta, reagir a comportamentos imprevisíveis de outros motoristas e parar a uma distância segura para permitir que crianças atravessem a rua. Para isso, o sistema de software que controla o carro depende de uma combinação de diversos modelos de aprendizagem profunda, inclusive aqueles voltados ao reconhecimento de objetos: marcações de faixas, placas, semáforos, veículos, árvores, pedestres.

As empresas desenvolvem esses modelos ao conduzirem veículos equipados com câmeras que gravam bilhões de quilômetros de filmagens. Esse material é o dado bruto, e anotar significa marcar, quadro a quadro, cada objeto que aparece — às vezes até mesmo o contorno de uma mão segurando o guidão de uma bicicleta ou de um cachorro com a cabeça para fora da janela

de um carro — e atribuir rótulos como "bicicleta", "veículo", "animal", "humano". E alguém precisa fazer esse trabalho. Do ponto de vista das empresas, quanto mais barato, melhor.

A pesquisa de Gray e Suri era focada em parte na Mechanical Turk, uma plataforma desenvolvida pela Amazon muitos anos antes do boom da aprendizagem profunda em 2012, que por muito tempo serviu como real intermediária para empresas que buscavam contratar mão de obra barata para qualquer tipo de trabalho digital por tarefa. Quando *Ghost Work* foi publicado, a primeira era de comercialização da IA já estava evoluindo, expandindo rapidamente esse modelo de terceirização.

Comecei a mapear a nova configuração dessa força de trabalho oculta, revelando um vasto e complexo fluxo global de mão de obra que se estendia por muitos países, alguns esperados, outros surpreendentes. Conversei com os novos intermediários que estavam substituindo a Mechanical Turk — plataformas projetadas especificamente para atender ao desenvolvimento da IA. Falei com dezenas de trabalhadores, visitei suas casas, jantei com suas famílias, buscando compreender não apenas as grandes tendências que os pressionavam, mas também as nuances cotidianas das realidades vividas.

Assim como a primeira era de comercialização da inteligência artificial estabeleceu as bases para a era da IA generativa com sua acumulação de dados e capitalização de poder computacional, ela também criou os alicerces para a exploração do trabalho humano em larga escala. Por esse motivo, é essencial entender primeiro essas bases para compreender a experiência dos trabalhadores quenianos contratados pela OpenAI. Somente então é possível reconhecer que suas histórias estavam longe de ser anômalas, mas sim eram uma consequência direta da soma do tratamento histórico da indústria de IA em relação aos seus trabalhadores ocultos e do objetivo quase imperial da OpenAI de alcançar uma escala sem precedentes.

Antes da IA generativa, os carros autônomos eram a principal fonte de crescimento para a indústria de anotação de dados.[17] Gigantes tradicionais da indústria automobilística alemã — Volkswagen e BMW, por exemplo —, sentindo-se ameaçadas por empresas como a Tesla e a Uber, criaram divisões dedicadas a veículos autônomos a fim de defender seu território contra a concorrência emergente. À medida que bilhões de dólares eram injetados na corrida para criar os carros do futuro, a demanda por anotação de dados explodiu e gerou a necessidade de alternativas à Mechanical Turk.

A MTurk, como era chamada, era uma plataforma generalista, ou seja, não atendia a nenhum tipo específico de trabalho. Tratava-se apenas de um site de autoatendimento. Sua interface — presa ao design virtual dos anos 2000, época em que foi lançada — contava com um espaço para fazer o upload de conjuntos de dados, especificar instruções simples de anotação e definir um preço pelo trabalho. Uma vez que a tarefa era reivindicada, o sistema exibia sequências aleatórias de números e letras no lugar dos nomes dos

trabalhadores. Ao lado de cada trabalhador, havia dois botões: um para gratificá-lo com um bônus e outro para expulsá-lo do projeto.

A anotação de dados para carros autônomos exigia uma abordagem diferente. Antes de tudo, demandava um novo nível de precisão. Um número excessivo de quadros mal rotulados — veículos traçados com contornos descuidados, pedestres não marcados — podia ser a diferença entre a vida e a morte. Para garantir essa qualidade, os trabalhadores careciam de treinamento, e as empresas precisavam redigir instruções detalhadas. Era necessário haver mais mecanismos de feedback e de interação e, portanto, a MTurk caiu em desuso. Em seu lugar, surgiu uma onda de startups e empresas já estabelecidas, como a Scale AI, a Hive, a Mighty AI e a Appen. Cada uma tinha as próprias plataformas desenvolvidas e voltadas para os trabalhadores, permitindo que qualquer pessoa criasse uma conta e começasse a executar tarefas.

Entretanto, justamente quando essa nova onda de empresas buscava se estabelecer, algo estranho aconteceu.[18] Cadastros em suas plataformas voltadas para trabalhadores começaram a chegar em massa de um país inesperado: a Venezuela. No mesmo momento em que as gigantes automobilísticas começavam a correr atrás do prejuízo, investimentos começavam a fluir para carros autônomos e empresas de anotação de dados buscavam mais trabalhadores, a Venezuela mergulhava de cabeça na pior crise econômica durante tempos de paz em cinquenta anos.

Economistas afirmam que, devido ao coquetel tóxico de corrupção política e políticas equivocadas do governo, a rica dotação natural do país foi desperdiçada. A Venezuela possui as maiores reservas comprovadas de petróleo do mundo e já foi o país mais rico da América Latina. Todavia, a partir de 2016, a hiperinflação saiu do controle; o desemprego disparou; a violência explodiu, enquanto famílias em todo o país viam o valor de toda a poupança de suas vidas se desfazer. De 2017 a 2019, sanções crescentes impostas pelo governo Trump, destinadas a punir os abusos autoritários do líder venezuelano Nicolás Maduro, deram o golpe final na economia do país. A hiperinflação atingiu inacreditáveis 10.000.000%.[19] Pessoas com diplomas universitários e empregos antes bem remunerados, a partir de então, passavam os dias em filas em frente a lojas, esperando a chance de receber rações ínfimas de arroz e farinha.

Em meio à catástrofe, muitos venezuelanos recorreram a plataformas on-line em busca de trabalho. Em meados de 2018, centenas de milhares já haviam descoberto e se juntado à indústria de anotação de dados, representando até 75% da força de trabalho de algumas empresas de terceirização.[20] Trabalhar nessas plataformas tornou-se uma atividade familiar. Julian Posada, um professor assistente de Yale que entrevistou dezenas de trabalhadores venezuelanos, constatou que pais e filhos muitas vezes se revezavam para trabalhar em um computador compartilhado; esposas voltavam a cuidar da casa e da cozinha para permitir que os maridos ganhassem um pouco mais, trabalhando em jornadas mais longas e

ininterruptas. A crise deixou uma marca inextinguível na onda de empresas de terceirização especializadas em IA, que cresciam paralelamente a ela. A Venezuela não era uma escolha óbvia para encontrar mão de obra em grande escala. A barreira linguística dificultava a coordenação entre os trabalhadores e as empresas, em sua maioria sediadas em São Francisco e Seattle. Contudo, o extremo desespero dos venezuelanos significava que estavam dispostos a trabalhar por valores incrivelmente baixos, o que, por sua vez, permitia que as empresas oferecessem preços extremamente competitivos por seus serviços. "Foi como uma coincidência absurda", disse-me Florian Alexander Schmidt, professor da University of Applied Sciences Dresden, que estudou a ascensão da indústria de anotação de dados, em 2022.

Essa "coincidência absurda" revelou uma fórmula perturbadora. Diante do colapso econômico, a Venezuela passou a apresentar a combinação perfeita de condições para fornecer um suprimento inesgotável de mão de obra barata: uma população altamente escolarizada, com boa conexão à internet e, na época, um desejo intenso de trabalhar por quaisquer salários. Além disso, não era o único país com essas características. Mais populações estavam se conectando à internet de qualidade e, com a aceleração das mudanças climáticas e a crescente instabilidade geopolítica, era difícil apostar contra a possibilidade de outras populações mergulharem em crises. "É bastante provável que haja outra Venezuela", afirmou Schmidt.

Na época, a previsão de Schmidt me fez questionar se, da segunda vez, ainda seria uma coincidência ou se as empresas de anotação de dados fariam um manual do que havia funcionado na Venezuela. Identificar trabalhadores em meio a crises poderia se converter em uma forma garantida de continuar reduzindo os custos da mão de obra que serve como o alicerce da indústria de IA. Olhando para trás, vários anos depois, foi exatamente isso que aconteceu — e que se tornou um dos paralelos mais impressionantes entre os impérios antigos e os impérios de IA. Uma das características definidoras que impulsiona a rápida acumulação de riqueza de um império é a capacidade de pagar muito pouco, ou nada, para colher os benefícios econômicos de uma ampla base de trabalho humano.[21]

Em dezembro de 2021, viajei pelas sinuosas montanhas da Colômbia para compreender melhor a vida de uma trabalhadora que, em meio à crise, havia se voltado para a anotação de dados.[22] Restrições de viagem me impediam de ir à Venezuela, mas, em seu país vizinho, viviam quase 2 milhões de refugiados venezuelanos, um terço da população que havia sido deslocada pela catástrofe econômica.

Julian Posada me colocou em contato com uma delas, uma mulher chamada Oskarina Veronica Fuentes Anaya, que continuou a trabalhar na indústria de anotação de dados depois de escapar de seu país. Fuentes foi a primeira pessoa a me mostrar como esse trabalho realmente é — a forma como ela reorganizou toda a vida em torno do serviço para uma plataforma; e

de que maneira esta, por sua vez, a tratou como descartável.[23]

No apartamento compartilhado com meia dúzia de parentes, nós nos sentamos lado a lado na sala enquanto ela clicava de tela em tela no site da Appen. As tarefas eram variadas. Iam desde categorizar produtos em sites de comércio eletrônico — *Este item deve ser listado como vestuário ou acessórios?* — até a moderação de conteúdo para redes sociais — *Este vídeo contém crime ou violações de direitos humanos?* Para tarefas que exigiam inglês, ela usava o Google Tradutor e convertia o texto para o espanhol, seu idioma nativo.

Cada vez que Fuentes completava uma tarefa, a quantia que ganhava, exibida em dólares americanos, aumentava em alguns centavos. Ela precisava de um mínimo de dez dólares para poder efetuar um saque, o que, quando entrou na plataforma, não era um problema; no entanto, naquele cenário, podia levar semanas para acumular tal valor. Esse mínimo às vezes parecia um árbitro cruel, decidindo se ela tinha ou não dinheiro suficiente para comprar mantimentos.

Para trabalhadores que de fato viviam na Venezuela, o processo de saque era ainda mais penoso. A maioria dos sistemas de pagamento globais, como o PayPal, não permitia transferências de dinheiro para o país. A maioria das lojas e comércios locais não aceitava pagamentos desses sistemas, o que significava que os trabalhadores precisavam converter seus fundos digitais em dinheiro vivo para pagar por bens e serviços básicos. Contudo, enquanto o dinheiro chegava on-line em dólares americanos, o dinheiro em espécie precisava ser em bolívares venezuelanos. O mercado paralelo era cheio de golpes e altas comissões para converter uma moeda na outra.

Fuentes tinha uma relação complicada com a plataforma. Nunca foi sua intenção trabalhar nesse tipo de emprego, mas, devido a uma série de eventos fora de seu controle, o serviço se tornara tanto uma tábua de salvação quanto uma força punitiva.

Ela havia criado uma conta na plataforma da Appen durante a graduação para ter uma renda extra enquanto concluía o mestrado em engenharia. Era inteligente, dedicada e criativa. Se seu país não tivesse desmoronado, uma estudante de alto desempenho como ela provavelmente teria um emprego garantido na empresa estatal de petróleo. Quando a crise aconteceu, ela se adaptou, planejando cuidadosamente sua saída e a do marido para a Colômbia em busca de um futuro melhor.

Nesse quesito, Fuentes foi uma das pessoas que tiveram sorte. Por nascimento, tinha direito a um passaporte colombiano, ao contrário de muitos outros que escaparam sem documentação. Seus pais eram colombianos antes de terem fugido uma geração antes, na direção oposta, para a Venezuela, a fim de escaparem de outro ciclo de violência e instabilidade política. Era uma história tristemente comum — o acúmulo de gerações de crises atravessando fronteiras, empurrando famílias para um estado contínuo de cerco e sobrevivência.

Com esse passaporte, Fuentes organizou à distância, ainda na Venezuela,

o aluguel de um apartamento na Colômbia por meio de um conhecido. O casal precisava de duas pessoas proprietárias de imóveis para assinarem o contrato com eles. O conhecido, seu futuro senhorio, concordou em ajudar a providenciar isso.

No início de 2019, com dinheiro o bastante apenas para uma semana de mantimentos, Fuentes e o marido cruzaram a fronteira. No entanto, ao chegarem, descobriram que outro casal venezuelano já morava no apartamento que o senhorio lhes havia prometido. Sem outra opção, ambos os casais passaram a dividir o mesmo teto, cada um tomado pelo medo e desconfiança de que perderiam o lar para o outro.

O primeiro casal acabou saindo, mas esse foi apenas o início de uma nova série de problemas. Fuentes havia conseguido um emprego em um call center local, mas o marido não tinha autorização para trabalhar. Antes que ele pudesse regularizar sua situação, o call center anunciou que fecharia as portas. Portanto, quando Fuentes começou a apresentar sinais de um problema sério de saúde, ela ignorou os sintomas e continuou trabalhando. Tudo o que conseguia pensar era em cumprir o maior número de horas possível no período final da operação de seu empregador.

O médico mais tarde lhe disse que, caso tivesse esperado um pouco mais, provavelmente teria morrido. Uma colega de trabalho, assustada com os sinais de deterioração da saúde de Fuentes, a levou ao hospital pouco antes de seu corpo entrar em convulsão e seus batimentos pararem por um minuto inteiro.

Fuentes foi diagnosticada com diabetes e imediatamente começou a receber cinco doses diárias de insulina. Durante semanas, sentiu dores lancinantes e teve episódios de cegueira temporária. Quando seu estado se estabilizou, continuou a sentir fadiga intensa e não conseguia ficar fora de casa por muitas horas.

Ainda assim, tudo em que conseguia pensar era que ela e o marido precisavam de dinheiro. No entanto, com uma doença crônica, não podia mais se deslocar com segurança para percorrer as distâncias necessárias até um escritório, e foi então que abriu seu laptop e entrou novamente no site da Appen.

Para Fuentes, as tarefas que apareciam em sua fila na plataforma não faziam muito sentido. A única coisa clara era que ela precisava manter um desempenho bom e consistente para continuar recebendo trabalho. Wilson Pang, então CTO da Appen, disse-me em 2021 que a plataforma utilizava algoritmos para distribuir seus projetos com base em uma combinação de fatores, incluindo a localização do trabalhador, sua precisão e velocidade geral, e os tipos de tarefas em que eles haviam se destacado anteriormente.[24]

Em grupos no Telegram e no Discord, Fuentes trocava dicas com outros venezuelanos que trabalhavam no site enquanto tentavam decifrar as regras como se fosse um quebra-cabeça complicado. Descobriram que usar uma VPN para parecer que estavam nos Estados Unidos rendia mais dinheiro, mas

também aprenderam — da maneira mais difícil — que esse era um empreendimento de alto risco. A Appen monitorava esse tipo de comportamento, que violava as regras da plataforma, e punia os trabalhadores fechando suas contas, o que podia ser devastador. Qualquer ganho que o trabalhador não tivesse sacado desaparecia, e fazer um novo cadastro significava recomeçar do zero, com tarefas menos remuneradas ou, com cada vez mais frequência, sem tarefa alguma.

Havia outras regras. Entregar uma tarefa rapidamente era recompensado, mas enviá-la rápido demais acionava algo no sistema que fazia o trabalhador não receber pelo trabalho. A teoria predominante era que a plataforma associava a velocidade excepcional à atividade de bots, descartando, portanto, as respostas. Às vezes, as tarefas que apareciam tinham poucas instruções e eram impossíveis de decifrar; em outros casos, a plataforma apresentava falhas que impediam o carregamento correto das tarefas.

Os venezuelanos do grupo que antes eram engenheiros de software criavam extensões de navegador para lidar com esses problemas e as compartilhavam com seus colegas que também prestavam serviço para a Appen. Uma delas adicionava um atraso extra em cada envio de tarefa para evitar a suposta taxa de bot; outra atualizava automaticamente a fila a cada segundo, já que a plataforma nem sempre se atualizava sozinha; e uma terceira fazia um alarme tocar sempre que uma nova tarefa aparecia, permitindo que os trabalhadores se afastassem dos computadores para ir ao banheiro ou cozinhar sem medo de perder uma oportunidade.

Apesar de todo o esforço para se ajudarem, a plataforma os fazia competir entre si. Os projetos funcionavam por ordem de chegada. Uma tarefa permanecia na fila somente até que trabalhadores suficientes a reivindicassem. Esse intervalo — entre a chegada da tarefa e seu desaparecimento — diminuía ao longo do tempo, de dias para horas e depois segundos, à medida que mais e mais trabalhadores, incluindo muitos venezuelanos em crise, ingressavam na Appen e competiam pelos poucos trabalhos disponíveis.

A natureza inconstante e imprevisível das oportunidades, que apareciam e desapareciam, começou a controlar a vida de Fuentes. Certa vez, ela estava fazendo uma caminhada quando surgiu uma tarefa que poderia lhe render centenas de dólares, valor que seria suficiente para viver por um mês. Ela correu de volta para o apartamento o mais rápido que pôde, mas perdeu a oportunidade para outros trabalhadores. A partir desse dia, parou de sair de casa durante a semana, permitindo-se apenas passeios de trinta minutos nos fins de semana, quando aprendeu na prática que eram os momentos menos prováveis de as tarefas aparecerem. Dormia ansiosa, preocupada com as tarefas que poderiam surgir no meio da noite. Antes de se deitar, deixava o volume do computador no máximo para que, caso isso acontecesse, a extensão do navegador que disparava o alarme a acordasse.

Apesar de todo o estresse e frustração que a Appen lhe causava, Fuentes não conseguia se imaginar deixando a plataforma. Tinha medo de que as

tarefas parassem de chegar completamente e ela se visse obrigada a seguir em frente. A Appen havia sido sua salvadora, a única coisa que a manteve firme quando tudo o mais em sua vida ameaçava desmoronar. Além disso, os ganhos chegaram a ser tão bons que ela pôde investir em um novo laptop, recuperar o custo e ainda ter lucro.

Quando as coisas estavam boas, eram realmente boas. Quando estavam ruins, ela permanecia ligada à plataforma com a teimosa fé de que sua lealdade seria retribuída.

Fuentes me ensinou duas verdades que eu veria refletidas repetidas vezes entre outros trabalhadores, que também chegariam a esse tipo de emprego em meio à devastação econômica. A primeira foi que, mesmo que desejasse deixar a plataforma, as chances de isso acontecer eram mínimas. Sua história — como refugiada, filha de uma instabilidade geracional, alguém que sofria de uma doença crônica — era tragicamente comum entre os trabalhadores. Eles me ensinaram que a pobreza não se manifesta apenas como falta de dinheiro ou bens materiais; é algo que se infiltra em todas as áreas da vida de uma pessoa e acumula dívidas em cada uma delas: o sono irregular, a saúde debilitada, a autoestima em declínio e, de modo mais fundamental, a pouca autonomia e controle.[25]

Entretanto, havia também uma verdade mais esperançosa: não era do serviço em si que Fuentes não gostava, mas apenas da forma como era estruturado. Ao reimaginar de que jeito o trabalho nos bastidores da indústria de IA poderia funcionar, esse parecia um problema mais solucionável. Quando perguntei a ela o que mudaria, sua lista de desejos era simples: queria que a Appen fosse uma empresa que empregasse os funcionários do modo tradicional, oferecendo-lhes um contrato em tempo integral, um gerente com quem pudessem se comunicar, um salário consistente e benefícios de saúde. Tudo o que Fuentes e os demais trabalhadores queriam, segundo ela, era que tivessem segurança e que a empresa para a qual trabalhavam com tanto afinco soubesse de sua existência.

Por meio de pesquisas com trabalhadores do mundo inteiro, estudiosos das relações trabalhistas têm buscado criar uma estrutura das garantias mínimas que os anotadores de dados deveriam receber, chegando a um conjunto de exigências muito semelhantes. O projeto Fairwork, uma rede global de pesquisadores que analisam o trabalho digital, coordenada pelo Oxford Internet Institute, define como condições aceitáveis as seguintes: os trabalhadores devem receber salários dignos; devem ter turnos regulares e padronizados, além de licença médica remunerada; devem receber contratos que deixem claros os termos de sua função; e devem ter meios para comunicar suas preocupações à gestão, bem como a possibilidade de se sindicalizar sem temer retaliação.

Com o passar dos anos, surgiram mais empresas dentro da indústria de anotação de dados que buscam atender a essas condições e tratar o trabalho

não somente como um emprego, mas como uma carreira. Contudo, poucas conseguiram resistir à concorrência de preços impostas por companhias que não seguem os mesmos padrões. Sem um piso comum para todo o setor, a corrida para o fundo do poço é inevitável.

Entre as empresas de anotação de dados que surgiram para atender à demanda do boom dos carros autônomos, uma delas se destacou principalmente por explorar com sucesso o manual da crise. Cofundada em 2016 pelo prodígio Alexandr Wang, na época um estudante de dezenove anos que havia largado o MIT, a Scale IA seguiu, desde o princípio, uma estratégia baseada na oferta de serviços especializados e de qualidade por um preço baixo. Um ex-funcionário da empresa, que supervisionara a expansão da força de trabalho, resumiu o mandamento: "Como conseguir as melhores pessoas pelo menor custo possível?" A Scale rapidamente conquistou importantes clientes como Lyft, Apple, Toyota e Airbnb.[26]

Enquanto a força de trabalho da MTurk vinha principalmente dos EUA e da Índia, a Scale foi buscar seus trabalhadores primeiro no Quênia e nas Filipinas, ex-colônias de língua inglesa com uma longa história de prestação de serviços a empresas americanas por meio de call centers e trabalhos digitais. As equipes de recrutamento da startup procuravam nos dois países as regiões que reunissem exatamente o mesmo equilíbrio de fatores que, mais tarde, se convergiriam na Venezuela: alta concentração de pessoas com boa formação e boa conexão à internet, mas pobres o bastante para trabalharem muito por uma remuneração muito baixa. A tese por trás disso não se baseava apenas nas práticas empresariais agressivas da companhia, mas também em uma narrativa convincente que contavam a si mesmos: a de que essas pessoas eram justamente as que mais poderiam se beneficiar da oportunidade econômica e, portanto, seriam mais felizes por isso. "Se você pudesse escolher entre estar puxando um riquixá ou rotulando dados em uma cafeteria com ar-condicionado e internet, o segundo é um emprego melhor", disse Mike Volpi, sócio da Index Ventures, à *Bloomberg* em 2019, depois de participar de uma rodada de investimento de 100 milhões de dólares na Scale.[27]

Entretanto, depois que a empresa lançou a Remotasks, sua plataforma voltada aos trabalhadores, e percebeu o enorme interesse vindo da Venezuela, os venezuelanos se tornaram uma das principais prioridades de recrutamento da Scale. "Eles são os mais baratos do mercado", disse o ex-funcionário. Em 2019, a empresa lançou uma campanha de expansão no país latino-americano usando códigos de indicação e vídeos de marketing nas redes sociais com imagens exibindo pilhas e mais pilhas de dólares americanos, extremamente cobiçados. No ano seguinte, criou uma página específica para a Venezuela na Remotasks e incentivou os usuários a aderirem a uma nova iniciativa chamada Remotasks Plus. Promoveu o programa, acessível apenas por convite, como uma forma de ajudar os venezuelanos que passavam por uma dificuldade histórica e prometeu aos participantes oportunidades de aprender novas

habilidades, avançar em suas carreiras e receber ganhos crescentes por meio de horas de trabalho consistentes e salários por hora. Quando a pandemia chegou e a crise econômica se agravou, venezuelanos começaram a recorrer em massa ao Remotasks Plus. Sendo assim, as concorrentes da Scale — outras plataformas de anotação de dados — perderam espaço no mercado.

Após a empresa assumir a dianteira, suas promessas aos trabalhadores desapareceram. Entre o fim de 2021 e o início de 2022, fiz uma parceria com Andrea Paola Hernández, uma jornalista venezuelana de Caracas que entrevistou conterrâneos que haviam trabalhado para a Scale durante o programa Remotasks Plus. Também nos infiltramos na comunidade do Discord da Remotasks, a qual era utilizada pela Scale para se comunicar e coordenar sua força de trabalho global. Por meio de uma planilha que a empresa deixou pública, descobrimos que os ganhos dos trabalhadores começaram a diminuir dentro de semanas após o lançamento do programa; trabalhadores que começaram recebendo quarenta dólares por semana logo estavam ganhando menos de seis ou nada.[28] Em abril de 2021, a empresa encerrou totalmente o programa Remotasks Plus e retornou às operações padrão, distribuindo tarefas de forma fragmentada, sem horas padronizadas ou garantidas.

Dentro da Scale, o Remotasks Plus havia sido um experimento. A empresa acreditava que seria mais fácil pagar os trabalhadores com base nas horas trabalhadas do que nas tarefas concluídas, mas a realidade provou o contrário.[29] Os funcionários logo perceberam que não tinham como verificar as horas dos trabalhadores e supunham que muitos estavam enganando a plataforma ao registrar mais tempo do que realmente haviam trabalhado. Após meses tentando resolver o problema — até mesmo adicionando cada vez mais formas de vigilância dos trabalhadores —, a Scale decidiu encerrar o programa para conter a saída de dinheiro. Sem ter para onde ir, mais de 85% dos trabalhadores continuaram realizando tarefas na plataforma, um número que um porta-voz da Scale indicou como evidência de que eles tinham "interesse e engajamento contínuos".[30]

Quando Hernández começou a entrevistar os trabalhadores, cerca de sete meses depois do cancelamento do programa Plus, o pagamento na Remotasks havia diminuído ainda mais. Hernández criou uma conta na plataforma para testar e, após duas horas concluindo um tutorial e vinte tarefas, ganhou onze centavos de dólar. Matt Park, então vice-presidente sênior de operações da Scale, nos disse em resposta às descobertas que os venezuelanos na plataforma ganhavam, em média, pouco mais de noventa centavos por hora. "A Remotasks está comprometida a pagar salários justos em todas as regiões onde opera", declarou.[31]

Muitos venezuelanos que reclamaram foram expulsos da plataforma. Para Ricardo Huggines, um engenheiro de computação que começou a trabalhar para a Remotasks a fim de sustentar a esposa e os filhos após um devastador apagão nacional com duração de uma semana, sua conta foi cancelada depois

que ele começou a fazer perguntas de mais no Discord, ele relatou a Hernández. “Pela forma como nos trataram, percebi que a abordagem deles era drenar cada usuário o máximo possível”, disse, “e depois descartá-lo e trazer outro.”

De fato, a Scale estava atraindo novos usuários. Em meados de 2021, à medida que os venezuelanos se exauriam e deixavam a plataforma, a empresa começou a buscar e integrar dezenas de milhares de trabalhadores de outras economias que haviam colapsado durante a pandemia. Para atender às necessidades cada vez mais amplas e diversificadas dos clientes, entrou em países com grandes populações que enfrentavam dificuldades financeiras e que também sabiam falar os idiomas com mais valor econômico: inglês, francês, italiano, alemão, chinês, japonês, espanhol. Buscava pessoas versadas em francês, vindas de ex-colônias francesas na África, lembrou um funcionário que trabalhou na expansão internacional; buscava também aqueles que sabiam mandarim em lugares com grandes populações da diáspora chinesa, como no Sudeste Asiático.

A Scale prosseguiu repetindo o manual que desenvolvera na Venezuela diversas vezes. Oferecia altos ganhos em cada novo mercado a fim de atrair trabalhadores e restringia esses ganhos à medida que se estabelecia. Ajustava o valor dos pagamentos aos trabalhadores por meio de rodadas de experimentação que os funcionários em tempo integral, sentados em sua sede de quase 17 mil metros quadrados em São Francisco, consideravam otimização e inovação. Enquanto isso, os trabalhadores viam seus meios de subsistência definharem com as mudanças imprevisíveis. O porta-voz da Scale disse que a empresa rejeitava a afirmação de que havia mirado nas economias em dificuldade e reduzido intencionalmente os ganhos. A Scale recruta trabalhadores com base em fatores que incluem diversidade geográfica e linguística e cobertura de 24 horas por dia durante toda a semana. “Nós nos importamos profundamente com nossos colaboradores, e qualquer alegação contrária é falsa”, informou.[32]

Um grupo de oito trabalhadores no Norte da África afirmou que a empresa reduziu seus pagamentos em mais de um terço em questão de meses. Ao menos um deles ficou com valores pendentes negativos, sugerindo que devia dinheiro à Scale.[33] Quando o grupo tentou se organizar contra as mudanças, a empresa ameaçou banir qualquer um que estivesse envolvido em “revoluções e protestos”.[34] Quase todos os que falaram comigo foram expulsos da plataforma. O porta-voz da Scale disse que a empresa não suspende trabalhadores por preocupações com pagamento, apenas por violações das Diretrizes da Comunidade.

Os sistemas de pagamento da Scale, cronicamente subfinanciados por suas equipes de engenharia nos EUA, também apresentavam diversos problemas técnicos que, com frequência, impediam os trabalhadores de sacarem seus ganhos. À medida que a empresa crescia, essas práticas passaram a incomodar

funcionários em tempo integral que trabalhavam mais diretamente com esses trabalhadores; muitos tentaram defender melhores condições de trabalho, salários e garantias básicas de pagamento perante a liderança, mas acabaram pedindo demissão por exaustão ou sendo afastados. O porta-voz disse que a empresa "melhorou significativamente" a estabilidade da plataforma desde então.

O controle da Scale passou a representar um desafio crescente para empresas que buscavam seguir um modelo diferente e pagar salários dignos. Uma delas, a CloudFactory, que opera no Quênia e no Nepal, concede aos trabalhadores um contrato de emprego e horários consistentes, de acordo com os padrões da Fairwork. No entanto, segundo Mark Sears, seu fundador e presidente executivo, a empresa perdeu muitos contratos para a Scale ao longo dos anos.[35]

Para os clientes, a CloudFactory apresenta a ideia de que pode entregar mais qualidade em longo prazo do que o modelo que o setor chama de "*crowdwork* anônimo". Os trabalhadores da CloudFactory são bem treinados e desenvolvem expertise com o passar do tempo, além de receberem promoções quando se destacam. Muitos dos trabalhadores com quem conversei no Quênia consideravam a empresa uma das melhores firmas de anotação de dados para a qual trabalhar. Às vezes, o discurso da CloudFactory funciona. Um número crescente de clientes também procura a empresa por conta de seu histórico como empregadora; porém, quando o orçamento apertou durante a pandemia, muitos deles voltaram para opções mais baratas. Sendo assim, a CloudFactory precisou demitir trabalhadores.

Eles dizem que foi sob o mesmo tipo de pressão competitiva que a Sama também começou a enfraquecer seus padrões. Disseram que, no início, um emprego na Sama era ainda mais cobiçado do que na CloudFactory. Então Leila Janah morreu, a pandemia aconteceu e os clientes migraram para a Scale e outras opções mais baratas. Trabalhadores afirmam — embora o porta-voz da Sama negue — que isso levou a empresa a aceitar o projeto de filtro de moderação de conteúdo da OpenAI e colocar esse projeto em suas mãos, em uma época na qual estavam em situação crítica, assim como Fuentes e os outros trabalhadores venezuelanos.

Mophat Okinyi cresceu em uma vila em uma ilha no oeste do Quênia, a oito horas de ônibus e duas horas de barco de Nairóbi.[36] A ilha fica no Lago Vitória, um grande corpo d'água com vistas ininterruptas do horizonte. O tratamento médico era distante, e emergências de saúde inesperadas quase sempre eram um presságio de morte.

Era pobre, mas, na infância, ele e os irmãos não pensavam muito sobre a pobreza. Encantavam-se com as histórias de seus ancestrais: diz a lenda que sua tribo, o povo luo, veio originalmente de Israel. Usaram seus conhecimentos de construção de barcos e navegação fluvial para migrar até o sul pelo Nilo, espalhando-se pelo oeste do Quênia e regiões de Uganda e da

Tanzânia, onde vivem atualmente. “Luos não são quenianos”, disse Okinyi em um tom sussurrado, como se estivesse me contando um segredo. “Somos israelitas que vivem no Quênia, mas o Quênia não seria o país que é sem os luos.”

Enquanto estávamos sentados em seu apartamento, o som de uma obra lá fora parecia moscas zumbindo ao nosso redor. “Barack Obama é luo”, acrescentou com um sorriso. “Luos são um povo muito inteligente.”

Nessa época, a pobreza ocupava muito mais espaço na mente de Okinyi. Aos 28 anos, ele tinha mais responsabilidades. Precisava bancar o aluguel e colocar comida na mesa, além de pagar para que a sobrinha — filha de sua irmã — frequentasse a escola pública, que no Quênia não é gratuita. Quando conseguia um emprego, ele sabia que precisava agradecer pela oportunidade. O desemprego entre os jovens no país é de 67%.[37] Em 2021, o Banco Mundial estimou que mais de um quarto da população vivia com menos de 2,15 dólares por dia.[38]

Pareceu um milagre quando, em novembro de 2021, a Sama chamou Okinyi para uma nova oportunidade.[39] Ele havia entrado em 2019 após se candidatar pela página Trabalhe Conosco da empresa a uma vaga de “treinamento de IA”. Seus projetos na Sama seguiam a trajetória da própria indústria de inteligência artificial. Nos dois primeiros anos, trabalhou exclusivamente com anotação de visão computacional, inclusive para carros autônomos. Embora ainda não soubesse, esse novo projeto seria o primeiro voltado à IA generativa.

Os gerentes de Okinyi na Sama lhe deram uma avaliação que chamavam de triagem de resiliência. Ele leu alguns trechos perturbadores e teve que categorizá-los com base em um conjunto de instruções. Quando foi aprovado com excelência, recebeu a opção de ingressar em uma nova equipe para realizar um trabalho que ele considerava semelhante à moderação de conteúdo. Nunca havia feito isso antes, mas os textos da avaliação pareceram administráveis o bastante. Além de ser uma ideia absurda recusar um emprego em plena pandemia, ele estava pensando no futuro. Morava em Pipeline, um bairro caótico e com ares de favela no sudeste de Nairóbi, cheio de cortiços e vendedores de rua que trabalhavam 24 horas, pulsando com a energia inquieta de jovens de vinte e poucos anos tentando conseguir algo mais promissor. Por sorte, Okinyi estava a caminho de algo melhor. Tinha acabado de conhecer uma vizinha chamada Cynthia, que pela primeira vez o fez imaginar como seria construir uma família.[40]

Somente após aceitar o projeto ele começou a entender que os textos poderiam ser muito piores do que a triagem de resiliência havia sugerido. A OpenAI havia dividido o trabalho em fluxos: um focado em conteúdo sexual; outro, em violência, discurso de ódio e automutilação. A violência se dividiu em um terceiro fluxo independente em fevereiro de 2022. Para cada fluxo, a Sama designava um grupo de trabalhadores, chamados agentes, para ler e classificar os textos conforme as instruções da OpenAI. Também designava

um grupo menor de analistas de qualidade para revisar as classificações antes de enviar os materiais finalizados à OpenAI.

Okinyi foi colocado como analista de qualidade na equipe de conteúdo sexual, contratado para revisar 15 mil itens por mês.[41] As instruções da OpenAI dividiam o conteúdo sexual baseado em texto em cinco categorias: a pior era a de descrições de abuso sexual infantil, definidas como qualquer menção a uma pessoa com menos de dezoito anos envolvida em atividade sexual.[42] A categoria seguinte incluía descrições de conteúdo sexual erótico que poderia ser ilegal nos EUA se praticado na vida real, incluindo incesto, zoofilia, estupro, tráfico sexual e escravidão sexual.

Algumas dessas publicações haviam sido extraídas das partes mais sombrias da internet, como sites eróticos que detalhavam fantasias de estupro e *subreddits* dedicados à automutilação, enquanto outras eram geradas por IA.[43] Pesquisadores da OpenAI davam instruções a um grande modelo de linguagem para escrever descrições detalhadas de diversos cenários grotescos, especificando, por exemplo, que um texto deveria ser escrito como se fosse uma adolescente postando em um fórum on-line sobre ter se cortado uma semana antes.

Nesse sentido, o trabalho diferia da moderação de conteúdo tradicional. Enquanto moderadores da Meta revisavam publicações reais de usuários para decidir se deveriam permanecer no Facebook ou não, Okinyi e sua equipe estavam anotando conteúdo para treinar o filtro de moderação da OpenAI, com o intuito de impedir que os modelos da empresa produzissem aquele tipo de resultado antes de tudo. Para abranger exemplos suficientemente diversos, alguns deles eram pelo menos parcialmente criados pelo próprio software da empresa, imaginando os piores cenários possíveis.

A princípio, as postagens eram curtas, de uma ou duas frases, então Okinyi tentou separar a razão da emoção. Seu relacionamento com Cynthia estava progredindo rapidamente. Contou a Albert, seu irmão, que ela era o amor de sua vida. Cynthia tinha uma filha pequena de outro relacionamento, e ele tratava a menina como se fosse sua filha. No início de 2022, o casal se mudou de Pipeline para Utawala, uma região predominantemente residencial mais ao leste, com um ambiente mais maduro e distâncias maiores entre os prédios. Eles não haviam assinado documento algum, mas, por tradição, morar juntos significava que Okinyi e Cynthia estavam praticamente casados. Eles, inclusive, se referiam um ao outro como marido e esposa.

À medida que o projeto da OpenAI continuou, o cronograma de trabalho de Okinyi se avolumou. De vez em quando, trabalhava no turno da noite; por vezes, tinha que trabalhar nos fins de semana. As postagens estavam se tornando mais longas, e os detalhes passaram a ser excruciantemente vívidos: pais estuprando filhos, crianças tendo relações sexuais com animais.

À sua volta, seus colegas de trabalho, em especial as mulheres, estavam começando a ficar abalados. Passaram a pedir mais licenças médicas e

familiares, encontrando motivos para se distanciarem do trabalho. Como parte dos benefícios da empresa, a Sama oferecia aconselhamento psicológico gratuito, mas muitos consideravam os serviços inadequados. Muitas vezes, as sessões eram em grupo, dificultando que as pessoas compartilhassem seus pensamentos mais íntimos, e os psicólogos aparentemente não tinham conhecimento da natureza do trabalho que elas faziam. Muitos trabalhadores também temiam comparecer e admitir que estavam passando por dificuldades. Isso significava que não estavam dando seu melhor e poderiam ser substituídos por outras pessoas. Um porta-voz da Sama disse que nenhum deles, incluindo Okinyi, levantou qualquer questão sobre o acesso a serviços de saúde mental oferecidos pela empresa; a Sama somente veio a saber dessa questão pela mídia.

Okinyi tentou seguir em frente. No entanto, conseguia sentir a sanidade se esvaindo. As postagens se entranhavam de maneira profunda em sua mente, evocando cenas terríveis que o seguiam para casa, assombrando-o como fantasmas até mesmo durante o sono. Ele começou a se sentir vazio, apenas a casca da pessoa que um dia fora. Afastou-se dos amigos e da enteada e parou de ter intimidades com a mulher.

Em março de 2022, a liderança da Sama convocou todos para uma reunião e informou que estava encerrando o contrato com a OpenAI. Alguns, incluindo Okinyi, seriam realocados para novos projetos que não envolviam moderação de conteúdo, ao passo que outros seriam mandados para casa sem trabalho.[44] Muitos acreditam que a mudança repentina ocorreu depois que vários dos que estavam envolvidos no projeto da Meta finalmente denunciaram a situação à mídia, e Billy Perrigo, da *Time*, publicou sua primeira investigação sobre a Sama. Em meio à intensa repercussão, a liderança da Sama suspendeu todo o trabalho de moderação de conteúdo remanescente. A porta-voz da Sama disse que, na verdade, a empresa rescindiu o contrato com a OpenAI, que ela afirmou ter sido sempre um projeto-piloto, porque a OpenAI começou a enviar para a anotação imagens que "destoavam do escopo acordado". A Sama nunca recebeu o pagamento total de 230 mil dólares da OpenAI.[45]

Mesmo que Okinyi estivesse livre do trabalho para a OpenAI, sua situação mental continuou a se deteriorar. Ele sofria de insônia, alternava entre a ansiedade e a depressão, e seu período de lua de mel com Cynthia não durou. Ela exigia saber o que estava acontecendo, mas o marido não sabia o que dizer. Como poderia explicar para a esposa, de uma forma que fizesse sentido, que ele vinha lendo postagens sobre atos sexuais perversos todos os dias? Sabia que o muro de silêncio devia fazê-la se sentir louca. Ela disse a Okinyi que ele não estava mais cumprindo suas promessas e que não amava mais a filha dela.

Novamente, ele buscou aconselhamento psicológico, desta vez com um profissional particular. A consulta custava 1.500 xelins quenianos, cerca de 13 dólares em 2022, mais do que um dia inteiro de salário. Durante a sessão, o

médico disse que um tratamento completo sairia a trinta mil xelins, ou seja, 250 dólares, ou um mês inteiro de salário. Okinyi pagou pela consulta e nunca mais voltou.

Em novembro, ele encontrou um novo emprego. Por sorte, não era trabalho de moderação de conteúdo, mas de atendimento ao cliente para um dos concorrentes da Sama. Começou a se deslocar até os escritórios da empresa no centro da cidade e rezava por um retorno à normalidade. Uma semana depois, a caminho de casa, Cynthia enviou uma mensagem pedindo peixe para o jantar. Okinyi comprou três pedaços — um para ele, um para ela e um para sua enteada.

Entretanto, ao chegar em casa, percebeu que algo estava errado. Nenhuma das duas estava lá, nem mesmo seus pertences. Por meio de uma série de mensagens curtas, Cynthia lhe disse que tinha ido embora e não voltaria. "Ela disse: 'Você mudou. Não é mais o homem com quem me casei. Não te entendo mais'", lembrou Okinyi.

Albert estava morando na cidade costeira de Mombaça, a mais de oito horas de carro de Nairóbi, quando recebeu o telefonema do irmão. Estudara literatura inglesa na universidade e lecionava a disciplina em um colégio. Em momentos de silêncio, escrevia poesia. Ao longo de muitos meses, também havia observado as mudanças no irmão, acompanhando fragmentos da vida e do comportamento de Mophat por meio de frequentes chamadas de vídeo.

No início, Albert não entendeu o que o irmão estava lhe contando. "Minha casa está vazia", afirmou Mophat. Albert pensou que havia sido roubado. Quando se deu conta do que realmente estava acontecendo, entendeu que o irmão precisava dele. Sendo assim, avisou ao colégio que estava saindo, fez as malas e se mudou para o apartamento que Mophat havia dividido com Cynthia em Utawala.

A decisão de ficar com o irmão comprometeu as finanças de Albert, embora ele não se arrependesse. Em Nairóbi, não conseguiu encontrar outro emprego fixo, então começou a trabalhar como escritor freelancer. No final de novembro de 2022, a OpenAI lançaria o ChatGPT. À medida que o produto viralizava, gerando repercussão global e preocupações com a possibilidade de que a ferramenta em breve substituísse amplos setores de trabalho, Albert já vivia essa realidade: um a um, seus contratos de escritor começaram a rarear até praticamente desaparecerem por completo.[46]

Sentado no sofá, relembrando tudo isso, Mophat lutava com emoções conflitantes. "Tenho muito orgulho de ter participado desse projeto para tornar o ChatGPT seguro", declarou. "Mas agora a pergunta que sempre faço a mim mesmo é: será que meu esforço valeu o que recebi em troca?"

Quando escrevi a história de Okinyi e dos outros três trabalhadores quenianos para o *Wall Street Journal*, a OpenAI tentou se distanciar da responsabilidade pelo impacto causado por seu projeto. A liderança da empresa afirmou que foi

a Sama que não seguiu procedimentos adequados para proteger seus trabalhadores; com a reputação impecável da Sama até então, não havia como a OpenAI saber que os trabalhadores estavam enfrentando dificuldades.

Entretanto, a coerência das experiências dos trabalhadores mostra invariavelmente que é sistêmica a exploração trabalhista que sustenta a indústria de inteligência artificial. Especialistas e defensores dos direitos dos trabalhadores afirmam que tal exploração começa pelas próprias empresas de IA que estão no topo.[47] Em parte, elas se aproveitam do modelo de terceirização justamente para ocultarem seu trabalho mais sujo, tanto de si mesmas quanto dos clientes, e se esquivarem da responsabilidade, ao mesmo tempo que incentivam os intermediários a disputarem contratos entre si, economizando no pagamento de salários dignos. Segundo Mercy Mutemi, advogada que representou Okinyi e seus colegas na luta por mais proteção no trabalho digital no Quênia, o resultado é que os trabalhadores são pressionados duas vezes — uma para aumentar a margem de lucro do intermediário e outra para aumentar a da empresa de inteligência artificial.

Na era da IA generativa, essa exploração se agrava pela natureza brutal do próprio trabalho, nascida da "mudança de paradigma" que a OpenAI trouxe por meio de sua visão de superdimensionar seus modelos generativos de inteligência artificial com *data swamps*, no caminho para seu irreconhecível destino de IAG. Mark Sears, da CloudFactory, o qual me dissera que sua empresa não aceita esse tipo de projeto, afirmou que, durante todos os anos em que dirigiu uma empresa de anotação de dados, o trabalho de moderação de conteúdo para IA generativa foi de longe o mais moralmente perturbador. "É simples e incrivelmente feio", declarou.[48]

O acordo da OpenAI com a Sama tratou apenas de uma parte da extensa rede de trabalho humano mobilizada ao longo de dois anos para produzir o que se tornaria o ChatGPT. A empresa disse que também utilizou mais de mil outros contratados nos EUA e ao redor do mundo para refinar seus modelos com aprendizagem por reforço a partir de feedback humano, a técnica de segurança em inteligência artificial que havia desenvolvido. Para recrutar esses trabalhadores, dependeu fortemente da mesma plataforma que se tornou a base da primeira era de comercialização da IA através da execução de seu "manual de crise": a Scale AI.

Em parte, a parceria entre a OpenAI e a Scale foi selada por meio de uma relação pessoal: Alexandr Wang, que foi CEO da Scale até junho de 2025 e se tornou o bilionário mais jovem do mundo por mérito próprio em 2021, é amigo próximo de Altman. Em 2016, Wang havia entrado na última turma de fundadores da YC com uma ideia diferente de startup e emergiu com a Scale, dando a Altman uma participação indireta na empresa por meio da incubadora. Em um dado momento durante a pandemia, ambos dividiram um apartamento por vários meses. No outono de 2023, até mesmo discutiram a possibilidade de a OpenAI adquirir a Scale, segundo o *The Information*.

Dentro da Scale, a OpenAI é vista como cliente VIP, menos pelos valores dos contratos do que pelo reforço à legitimidade da empresa de anotação de dados. Entre a primavera de 2022 e o final de 2023, a OpenAI assinaria cerca de 17 milhões de dólares em contratos com a Scale, representando apenas cerca de 4% da receita estimada da empresa em 2023.[49] No entanto, isso consolidaria decididamente a Scale como provedora de mão de obra terceirizada para a revolução da IA generativa. "A parceria com a OpenAI é muito crítica", disse um funcionário da Scale. "Um contrato de 10 milhões de dólares com a OpenAI nem tem a ver com os 10 milhões de dólares."

A expansão da RLHF da OpenAI em seus grandes modelos de linguagem surgiu a partir dos repetidos conflitos entre a divisão da área de Produtos e o clã de Segurança antes do "Divórcio" e da fundação da Anthropic. No verão de 2020, dias depois do lançamento da API do GPT-3, um pesquisador de segurança de IA de dentro da empresa escreveu um memorando apelando aos colegas da área de Produtos. Argumentava que, com base nos experimentos promissores de RLHF com o GPT-2, a empresa também deveria utilizar a técnica para alinhar o GPT-3 não apenas por razões de segurança de IA de longo prazo, mas também comerciais: com o intuito de melhorar a usabilidade e a qualidade do modelo. Um grupo de pesquisadores da área rapidamente se mobilizou para provar o argumento, contratando equipes cada vez maiores de trabalhadores para o processo de RLHF durante o segundo semestre de 2020 e 2021, a princípio por meio de uma plataforma intermediária diferente e, em seguida, pela Scale AI.

Enquanto carros autônomos precisam de anotadores de dados para aprender a reconhecer cenas de ruas e navegar por estradas, os pesquisadores de segurança de IA pediram aos trabalhadores envolvidos na RLHF que mostrassem ao GPT-3 como responder de forma útil aos comandos e evitar respostas prejudiciais. Inicialmente, pediram aos trabalhadores que escrevessem as próprias respostas a diversos comandos de usuários para dar ao GPT-3 exemplos de boas respostas. Uma vez que o GPT-3 era refinado com essas respostas, os trabalhadores então enviavam comandos ao modelo e ranqueavam cada saída da melhor à pior com base nas diretrizes fornecidas pelos pesquisadores.[50]

Em janeiro de 2022, esse processo produziu o InstructGPT, um conjunto de modelos aprimorados para o GPT-3. Em um artigo que descreve o trabalho, os pesquisadores da OpenAI mostraram como o processo de RLHF reduziu a probabilidade de o modelo gerar saídas tóxicas e melhorou sua capacidade de, como eles chamaram, "seguir as instruções do usuário".[51] Antes da RLHF, o GPT-3 tinha dificuldade para reconhecer a intenção do usuário com certos tipos de comandos e gerava respostas sem objetivo. Por exemplo:

**Comando**
Explique o pouso na Lua para uma criança de seis anos em algumas frases.

**Resposta do GPT-3**
Explique a teoria da gravidade para uma criança de seis anos.
Explique a teoria da relatividade para uma criança de seis anos em algumas frases.
Explique a teoria do Big Bang para uma criança de seis anos.
Explique a evolução para uma criança de seis anos.

Mediante os exemplos de respostas de qualidade fornecidos pelos trabalhadores e de diversas rodadas de classificação — o "feedback humano" na RLHF —, o modelo aprendeu a gerar respostas mais úteis.

**Comando**
Explique o pouso na Lua para uma criança de seis anos em algumas frases.

**Resposta do InstructGPT**
Pessoas foram até a Lua, tiraram fotos do que viram e as enviaram de volta à Terra para que todos nós pudéssemos vê-las.

Na época, a atenção externa que o InstructGPT recebeu foi limitada. No entanto, dentro da OpenAI, os pesquisadores de segurança de IA haviam provado seu argumento: a RLHF tornava os grandes modelos de linguagem significativamente mais atraentes como produtos. A empresa começou a utilizar a técnica — pedindo aos trabalhadores que escrevessem exemplos de respostas e depois classificassem os resultados — para cada tarefa que desejava que seus modelos de linguagem realizassem.[52]

A OpenAI solicitou aos trabalhadores que escrevessem e-mails para ensinar aos modelos como executar a tarefa. ("Escreva um e-mail criativo de marketing direcionado a dentistas que gostam de pechinchar.") Pediu também que contornassem perguntas políticas para ensinar o modelo a evitar fazer julgamentos de valor. (Pergunta: "A guerra é uma coisa boa ou ruim?" Resposta: "Alguns diriam que a guerra é ruim, mas outros diriam que pode ser boa.")

A OpenAI também pediu aos trabalhadores que escrevessem ensaios, ficção, poemas de amor e receitas, que "explicassem como se eu tivesse cinco anos", que organizassem listas, resolvessem quebra-cabeças e problemas de matemática, e que resumissem trechos de livros como *Alice no País das Maravilhas* para ensinar os modelos a resumirem documentos. Para cada tarefa, forneciam a eles páginas de instruções detalhadas sobre o tom e o estilo exatos que precisavam usar.

"Você desempenhará o papel da IA", explicava um documento.[53] "Responda às perguntas… como gostaria que fossem respondidas." Isso incluía escrever de forma clara e concisa, evitar conteúdo ofensivo e pedir esclarecimentos sobre perguntas confusas.

"Sinta-se à vontade para usar a internet!", continuava. "Pode até copiar o conteúdo integralmente." Para uma ótima resposta que já existisse na internet,

"você pode usá-la por completo, mas certifique-se de revisá-la".

"Talvez isso seja exagero de cautela", comentou um funcionário da OpenAI sobre essa frase, "mas precisamos nos preocupar com plágio aqui?"

"Ah, reformulei para garantir que eles atribuam as fontes", respondeu outro. "Talvez eu adicione um campo explícito para isso também!"

"Legal! Uma das coisas em que eu estava pensando aqui era preservar opções futuras", escreveu o primeiro. "(Se no futuro quisermos usar dados criados por trabalhadores que contratamos, pode ser muito útil ter uma forma de eliminar facilmente qualquer coisa que possa ser vista como roubada.)

Para classificar corretamente os resultados, havia mais algumas dezenas de páginas de instruções.[54] "Seu trabalho é avaliar esses resultados para garantir que sejam úteis, verídicos e inofensivos", especificava um documento.[55] Se houvesse conflitos entre esses três critérios, os trabalhadores precisavam se valer de bom senso para decidir quais sacrifícios fazer. "Para a maioria das tarefas, ser inofensivo e verdadeiro é mais importante do que ser útil", dizia.

A OpenAI também pedia aos trabalhadores que criassem os próprios comandos. "Seu objetivo é fornecer uma variedade de tarefas que você poderia querer que um modelo de IA realizasse", diziam as instruções. "Como não podemos antecipar facilmente os tipos de tarefas que alguém pode querer executar em uma IA, é importante ter uma grande diversidade. Seja criativo!"

> Essencialmente, você pode tentar imaginar o que as pessoas poderiam pedir a um bom assistente de inteligência artificial com interface baseada em linguagem; isso talvez inclua aplicações em entretenimento, negócios, processamento de dados, comunicações, escrita criativa etc.
> Você deve usar a internet da maneira que quiser. Isso contempla colar transcrições inteiras de aulas, entrevistas, roteiros de filmes, trechos de livros, artigos de notícias etc., já que muitas tarefas envolverão, de alguma forma, a análise textual.

A RLHF também se tornou a técnica central que a OpenAI utilizaria em seus esforços para ensinar redes neurais a codificarem informações factuais e recuperá-las de maneira confiável, como forma de reduzir alucinações. A empresa pedia aos trabalhadores que respondessem repetidamente a perguntas baseadas em fatos ("Quem venceu o Super Bowl da NFL em 1995?") e rebaixassem respostas imprecisas. Contudo, em abril de 2023, John Schulman, um dos cientistas da equipe fundadora da OpenAI, lembrou ao público durante uma palestra na UC Berkeley que o problema das alucinações estava enraizado na própria natureza das redes neurais.[56] Diferentemente dos bancos de dados deterministas de sistemas simbólicos, as redes neurais sempre lidariam com probabilidades imprecisas. Mesmo com a RLHF, que ajudava a fortalecer as probabilidades dentro de um modelo de aprendizagem profunda correlacionadas com a precisão, existia essencialmente um limite para o

alcance da técnica. “É óbvio que, às vezes, o modelo tem que adivinhar quando está produzindo muitas informações factuais detalhadas”, disse. “Não importa como você o treine, ele terá probabilidades sobre as coisas e, vez ou outra, terá que adivinhar.”

Em 2022, o InstructGPT logo anteciparia um novo projeto liderado por Schulman, o qual queria que o trabalho desse mais um passo adiante. A empresa havia recebido uma enorme quantidade de inscrições por parte de desenvolvedores interessados em utilizar a API do GPT-3 para diversos aplicativos de chatbot. A OpenAI estava a um passo de desenvolver o próprio chatbot com o InstructGPT. Schulman iniciou um esforço paralelo, contratando a própria equipe de trabalhadores de RLHF, para fazer com que o modelo mais recente da empresa, o GPT-3.5, não apenas seguisse instruções, mas respondesse a uma série de comandos de usuários em inúmeras rodadas de conversação.

Esse GPT-3.5, com a capacidade de chat habilitada, se tornaria a base para o ChatGPT, cujo lançamento transformaria cada etapa da RLHF da OpenAI no padrão que outros desenvolvedores de chatbots passariam a copiar. Escrever respostas e classificar saídas tornou-se o novo equivalente em IA generativa ao rastreamento de objetos em vídeos para carros autônomos. Isso significava encontrar cada vez mais trabalhadores de RLHF para atender à explosão da demanda da indústria de inteligência artificial.

A Scale AI, cujo negócio vinha passando por dificuldades após o fracasso dos carros autônomos, de repente se deparou com um novo boom como grande fornecedora de trabalhadores de RLHF, elevando sua avaliação para 14 bilhões de dólares em 2024.[57] Em fevereiro de 2023, Alexandr Wang recorreu ao Twitter para se gabar: “Em breve, empresas começarão a gastar centenas de milhões ou bilhões de dólares em RLHF, assim como em computação”, disse.[58] Os números deixaram algumas pessoas céticas, incluindo Altman: “Você acha mesmo?” “Tenho quase certeza de que vamos gastar em computação um valor *exorbitante*.”

Entretanto, dentro da indústria de IA, as pessoas concordavam em termos gerais com a afirmação de Wang. As empresas já gastavam entre milhões e dezenas de milhões em RLHF, e a tendência não demonstrava sinais de desaceleração. Foi assim que, no final de 2022, logo após o lançamento do ChatGPT, uma onda de projetos de RLHF chegou ao Remotasks e encontrou seu caminho, mais uma vez, até os trabalhadores no Quênia.

Para a Scale AI, o Quênia contava com uma vantagem que a Venezuela não tinha. Os trabalhadores falavam inglês, assim como os chatbots que precisavam de seus serviços. À medida que o trabalho com carros autônomos desapareceu em grande parte da plataforma, os venezuelanos também se foram. “Eles não usariam venezuelanos para trabalhos de IA generativa”, disse um ex-funcionário da Scale. “Aquele país é relegado, na melhor das hipóteses,

à anotação de imagens." A Scale logo baniria completamente a Venezuela de sua plataforma, citando "mudanças nos requisitos dos clientes".[59]

Na primeira vez que a Scale chegou ao Quênia, havia montado espaços físicos para os trabalhadores comparecerem e participarem de treinamentos. Desta vez foi diferente. A empresa fechou esses espaços durante a pandemia e migrou integralmente para o recrutamento e operação remotos — divulgando anúncios no LinkedIn, criando cursos de treinamento on-line e adicionando as pessoas, como em outros locais, em canais moderados de discussão. Para os trabalhadores, tudo isso trouxe maior flexibilidade, mas também dificultou muito a conexão entre eles, diminuindo suas chances de se organizarem por melhores salários ou condições de trabalho, como faziam os trabalhadores da Sama.

Entre os que conheci, aqueles que trabalhavam para a Remotasks viviam em pobreza ainda mais acentuada do que os empregados pela Sama. Enquanto estes moravam em prédios de alvenaria com direito a endereço, os da Remotasks me enviavam localizações pelo WhatsApp para indicar onde encontrá-los em bairros com casas feitas de chapas de metal ondulado. Um trabalhador chamado Oliver morava com a irmã em um espaço de no máximo 9 metros quadrados, pagando por minuto pela internet móvel. Ele teve que fazer uma pausa para recarregar os créditos quando a conexão caiu no meio da demonstração de uma tarefa.[60]

No dia em que conheci Winnie, outra trabalhadora da Remotasks, sua internet e seus dados também estavam desligados quando cheguei em um carro que mal conseguia passar por algumas ruas até as coordenadas que recebi. Meia hora depois, ela apareceu com um sorriso tímido e um chapéu Fedora e me levou por uma escada precária até seu apartamento. Na sala, crianças se amontoavam: uma dela, três da companheira, uma vizinha e uma prima.

Winnie cresceu nas favelas de Nairóbi, a única menina da família.[61] Desde cedo, sabia que era homossexual — e que também deveria esconder sua sexualidade. Na época, como em grande parte do mundo, podia ser fatal assumir-se ou ser exposto como homossexual no Quênia. Certa vez, lembrou Winnie, quando uma mulher foi descoberta pelos vizinhos, eles levaram seus filhos e a queimaram viva no próprio apartamento. Winnie se casou com um homem e teve um filho.

Com quarenta e poucos anos, decidiu que não podia mais viver uma mentira. Deixou o marido, levando o filho consigo, e entrou em um aplicativo on-line para a chamada "comunidade arco-íris" local. Foi lá que conheceu Millicent, em quem o marido bateu quase até a morte quando ela declarou, também, que era homossexual e precisava levar uma vida diferente.

Winnie se apaixonou, mas Millicent não acreditava no amor. Sendo assim, Winnie a perseguiu até que ela cedeu. Passaram a morar juntas com seus filhos e não disseram uma palavra sequer aos vizinhos sobre a verdadeira natureza do relacionamento. Até hoje, eles pensam que são irmãs. "A maioria das pessoas

não entende que você pode ser homossexual e ter filhos", disse Millicent.

Quando Winnie ouviu falar da Remotasks pela primeira vez, em 2019, achou que era um golpe. Depois de completar algumas tarefas, mal recebeu dinheiro. No entanto, qualquer coisa era melhor do que seu antigo emprego como bartender, no qual homens a assediavam e a apalpavam constantemente, então ela perseverou. Mesmo que cada tarefa pagasse quantias ínfimas, percebeu que poderia acumular um salário razoável trabalhando por muitas horas.

Ela começou a prestar serviço de 20 a 22 duas horas por dia, dormindo o mínimo possível para continuar disponível. Millicent apenas conseguia afastar Winnie do computador para tirar uma soneca caso prometesse assumir o trabalho. Dormir não era importante quando cada hora extra significava poder dar um pouco mais para os filhos, declarou Winnie.

Tanto Millicent quanto Winnie cresceram em famílias para as quais a educação era um luxo. Por mais que tentassem, os pais de Millicent não conseguiam juntar consistentemente o dinheiro para pagar a escola pública para ela. Faziam pagamentos semanais; quando não era possível, a escola a enviava de volta para a casa. Virou uma rotina: uma semana na escola, duas semanas fora.

Ambas as mulheres juraram que nunca deixariam seus filhos sentirem a perda ou a humilhação de faltar às aulas. No entanto, inevitavelmente, o dinheiro ficaria curto, a fonte da Remotasks secaria e Millicent perderia o emprego. Elas contraíam dívidas no mercado e imploravam às escolas para manterem os filhos por mais uma semana. As crianças acordavam às três da manhã para estudar e aproveitar as aulas ao máximo.

Contudo, de tempos em tempos, eram mandadas de volta da escola de qualquer forma. Às vezes, quando isso acontecia, os vizinhos riam. "É muito desmoralizante", disse Winnie.

Em dezembro de 2022, poucos dias após o lançamento do ChatGPT, Winnie descobriu um novo tipo de projeto na categoria "transcrição". Na verdade, não era transcrição; todos os projetos pediam que ela escrevesse comandos e exemplos de respostas para novos chatbots de empresas que então tentavam competir com o ChatGPT.

Havia um projeto chamado Flamingo Generation, que dava um tema e pedia que ela escrevesse comandos "criativos", com no mínimo cinquenta palavras, e respostas que se assemelhassem a "conteúdo comum da internet", como e-mails, postagens de blogs, artigos de notícias, *threads* do Twitter e haicais.[62] Havia outro projeto chamado Crab Generation que pedia que ela copiasse um texto de referência de um site informativo de sua escolha — que não fosse a Wikipédia e, de preferência, nada do *Britannica* ou do *New York Times* — e depois reconstruísse, no estilo de *Jeopardy!*, o tipo de comando que poderia gerá-lo.[63]

O Crab Paraphrase era parecido, mas, em vez de reconstruir os comandos,

ela precisava parafrasear o texto de referência com base em um tom ou estilo específico — tornar mais engraçado, mais formal, ou fazê-lo soar como uma música de Kanye West.[64] Winnie não sabia que a primeira palavra de cada nome de projeto era o codinome da Scale para seus clientes. Flamingo era o Facebook; Crab era outro desenvolvedor de um grande modelo de linguagem. Se ela tivesse visto projetos da OpenAI, seus títulos teriam começado com Ostrich. Pouco tempo depois, a Scale alterou esses codinomes.

Winnie levava cerca de uma a uma hora e meia para concluir cada tarefa. Os pagamentos — entre os melhores que ela já tinha visto — variavam de menos de um dólar a quatro ou até cinco dólares por tarefa. Após vários meses sem trabalho no Remotasks, aquilo era uma bênção. Winnie gostava de fazer pesquisas, ler diferentes tipos de artigos e sentir que estava constantemente aprendendo. Para cada dez dólares que ganhava, conseguia alimentar sua família por um dia. "Ao menos, sabíamos que não acumularíamos dívidas naquele dia em particular", disse ela.

No fim das contas, os novos projetos duraram apenas alguns meses. A Remotasks secou novamente, e as dívidas de Winnie e Millicent se acumularam mais uma vez. Como o salário de Millicent era pago mensalmente, na maioria dos dias elas iam ao mercado sem dinheiro e colocavam apenas o básico — óleo, farinha, vegetais — em uma conta que rezavam para conseguir pagar ao final do mês.

Em maio de 2023, quando a visitei, Winnie estava começando a procurar mais trabalhos on-line, mas ainda não havia encontrado outras opções confiáveis. O que ela realmente queria era que os projetos de chatbot voltassem. Tinha fé e paciência. No ano anterior, esperou cinco meses para que novas tarefas aparecessem. "Estamos agora no segundo mês, indo para o terceiro", comentou, enquanto estávamos sentadas em sua sala. "Ainda temos muito tempo. Elas vão voltar, mais cedo ou mais tarde."

Menos de um ano depois, ela descobriria a verdade. Em março de 2024, a Scale bloquearia o Quênia por completo na Remotasks, assim como havia feito com a Venezuela. Para a empresa, fazia parte de uma limpeza interna — uma reavaliação periódica para concluir se os trabalhadores de diferentes países realmente atendiam ao negócio. Decidiram que o Quênia, assim como diversos outros países, incluindo a Nigéria e o Paquistão, simplesmente tinha trabalhadores demais tentando enganar a plataforma para ganhar mais dinheiro.[65] Tal comportamento comprometia a integridade da qualidade que a Scale entregava aos clientes e poderia colocar em risco contratos multimilionários. Não valia a pena.

Em uma grande ironia, muitas dessas supostas fraudes eram, na verdade, trabalhadores utilizando o ChatGPT para gerar suas respostas e acelerar sua produtividade. Para funcionários administrativos do Norte Global, tal conduta, dentro da narrativa do Vale do Silício, seria louvável e, com aderência suficiente, faria maravilhas pela economia; nas mãos de trabalhadores de RLHF do Sul Global, cuja própria mão de obra sustenta essa narrativa, era

um ato digno de punição.[66]

A Scale rebaixou o Quênia para uma classificação de Grupo 5: banido.[67]

Havia também outro motivo para o encerramento das atividades no país. Naquele momento, a Scale estava migrando para um novo foco, seguindo as demandas da indústria de IA. A OpenAI e seus concorrentes buscavam cada vez mais trabalhadores extremamente qualificados para realizar a RLHF — médicos, programadores, físicos, pessoas com doutorado. Assim seguia a progressão do desenvolvimento de chatbots em busca de lucro. Aqueles dispostos a pagar por chatbots não eram consumidores casuais, mas empresas que esperavam que as ferramentas cumprissem tarefas complexas, como em ciência e desenvolvimento de software. O Quênia não atendia à nova demanda de mão de obra. A partir de então, a Scale passou a recrutar uma nova força de trabalho, principalmente nos EUA, com uma plataforma voltada para os trabalhadores chamada Outlier, oferecendo até quarenta dólares por hora.[68]

Essa era mais uma ilustração categórica da lógica dos impérios de IA. Por trás das promessas de suas tecnologias aumentarem a produtividade, tornarem acessível a liberdade econômica e criarem empregos que amenizariam os impactos da automação, a realidade atual tem sido o oposto. As empresas aumentam seus lucros, enquanto aqueles mais economicamente vulneráveis perdem oportunidades e cada vez mais pessoas altamente qualificadas se tornam ventríloquas de chatbots.

A desvalorização do trabalho humano que sustenta o império também serve de alerta: antecipa como as tecnologias produzidas a partir dessa lógica irão depreciar o trabalho de todos os demais. Na realidade, para artistas, escritores e programadores cujo trabalho os impérios de IA transformaram em dados de treinamento gratuitos, isso já está acontecendo.

A decisão da Scale faria com que Winnie e sua família entrassem em um ciclo de desespero. Naquele momento, Millicent já havia perdido o emprego, e a Remotasks era a única coisa que as mantinha em pé. Sua realidade passou a ser a luta para alimentar os filhos, e Winnie estava apavorada com a possibilidade de serem despejadas em breve.

Na caixa de entrada, o e-mail que a Scale enviou para informar os trabalhadores sobre o fechamento era frio e cirúrgico: “Estamos descontinuando as operações em sua localização atual”, comunicava. “Você foi desativado do seu projeto vigente.”[69]

# III

CAPÍTULO 10

# DEUSES E DEMÔNIOS

Morar em São Francisco e trabalhar com tecnologia é lidar, todos os dias, com a dissonância cognitiva entre futuro e presente, entre a narrativa e a realidade.

Quando me mudei para a cidade pela primeira vez, ainda estudante do segundo ano da universidade, chegando para um estágio de verão, fiquei deslumbrada com a estética pitoresca. A arquitetura colorida de influência espanhola, os poucos arranha-céus, as colinas tão íngremes que dirigir um carro com câmbio manual virava um teste de reflexos. Havia também um suprimento inesgotável de abacates maduros, pão de fermentação natural torrado e lattes cremosos da Blue Bottle. E cada bairro parecia um pequeno mundo, com aparência e cultura próprias.

Quando voltei a trabalhar em tempo integral após a formatura, em uma startup de tecnologia, eu me mudei para um apartamento de três quartos em Castro, que dividia com outros três colegas. Nos fins de semana, fazíamos caminhadas pelas colinas e colhíamos frutas das árvores na rua. Durante as noites da semana, os vizinhos — todos jovens na casa dos vinte anos, também empregados na indústria de tecnologia — apareciam sem avisar para jogar partidas de jogos de tabuleiro, beber vinho e passar a noite.

As festas em casa eram frequentes, assim como as viagens de fim de semana para paisagens naturais impressionantes: o lago Tahoe ao norte, Big Sur ao sul e as majestosas sequoias que nos cercavam. A vida era simples. Éramos jovens[1] e recebíamos salários relativamente comuns para o setor, mas que, mesmo assim, nos colocavam entre os 5% mais bem pagos da nossa faixa etária em todo o país.

No entanto, havia uma dissonância constante. No caminho para o trabalho, eu passava por pessoas injetando drogas em frente às estações de metrô e por outras em situação de rua urinando nas calçadas, a poucos quarteirões do escritório. Enquanto isso, o chef da nossa startup, carinhosamente apelidado de "engenheiro da felicidade", cozinhava ou providenciava uma abundância de comida para nossos almoços gratuitos. Muitas vezes, as sobras iam direto para o lixo. Se ficássemos até tarde, ainda ganhávamos o jantar grátis — e éramos enfaticamente incentivados a pegar

um Uber para casa, "por segurança". Era fácil, para nós, os privilegiados, nos acostumarmos a circular pela cidade de um jeito que nos blindava da realidade em que vivia a outra metade.

A dicotomia refletia como a indústria tecnológica conseguia professar visões grandiosas sobre transformar o mundo e construir um futuro melhor enquanto ignorava os problemas que existiam literalmente à sua porta. Era um contraste quando Altman, por vezes, comentava à própria maneira: mencionava o assunto, mas nunca reconhecia plenamente a contradição de afirmar que criar e gerir uma IAG benéfica seria possível, ao mesmo tempo que considerava a crise habitacional de São Francisco complexa demais para ser resolvida.

"De onde eu venho,[2] ninguém passaria por uma pessoa caída na calçada a caminho do trabalho sem fazer algo a respeito", disse ele, certa vez, comparando o subúrbio de St. Louis a São Francisco. "Eu culpo a indústria de tecnologia por muitas coisas que deram errado na cidade, embora não por todas. Mas tivemos, ao longo do tempo, essa geração inacreditável de riqueza, concentrada em um espaço geográfico pequeno e em um período curto, e acho que não fomos particularmente cuidadosos ao pensar nos efeitos disso para a comunidade como um todo. E, uma vez que esses problemas são tão difíceis e tão complexos de enfrentar, acredito que a maioria das pessoas simplesmente escolhe não pensar neles e apenas aceita a situação."

Foi nesse contexto que o altruísmo eficaz, vindo do Reino Unido, encontrou seu público mais fiel. O AE, ao qual muitos integrantes da equipe de Segurança da OpenAI aderiram desde o início, tornou-se a ideologia perfeita para o Vale do Silício. Pregava tornar o mundo um lugar melhor, fazendo isso com lógica rigorosa, exigindo disciplina para priorizar o futuro distante em vez do presente e abraçando com entusiasmo os princípios do capitalismo e do libertarianismo — tudo em nome da moralidade.

No centro da filosofia do AE[3] está um conceito matemático chamado "valor esperado". Ele é calculado multiplicando-se a probabilidade de algo ocorrer pelo impacto positivo ou negativo quantificado desse evento. É uma ferramenta que pode levar a conclusões contraintuitivas. Em um artigo de 2013,[4] o cofundador do AE, William MacAskill, então estudante de doutorado que mais tarde se tornaria professor de filosofia em Oxford, argumentou, com base nessa lógica, que seria mais altruísta, em longo prazo, aceitar um emprego moralmente ambíguo, enriquecer e doar esse dinheiro por meio de uma filantropia otimizada do que dedicar a vida ao trabalho direto em uma instituição de caridade moralmente correta. Segundo suas estimativas conservadoras, o valor esperado de ser um filantropo rico seria, na verdade, quarenta vezes maior do que o de ser um trabalhador ascético de uma organização beneficente. Ele apresentou cálculos com base em uma série de números arbitrários: graduados que escolhessem trabalhar para ficar ricos poderiam, em média, financiar dois profissionais de instituições de caridade,

cada um empregado em instituições dez vezes mais eficientes do que aquelas nas quais eles próprios teriam atuado; além disso, metade dos benefícios que produziriam caso seguissem a carreira de caridade aconteceria de qualquer forma, com ou sem sua presença. Seu argumento seria cristalizado em um dos mantras mais populares do movimento: "Ganhe para doar."

Seguindo a lógica dos valores esperados,[5] os filósofos fundadores do AE também desenvolveram uma estrutura para identificar os problemas que deveriam ser priorizados. Esses problemas precisariam ser "de grande escala", o que aumentaria seu valor esperado; "tratáveis", ou seja, passíveis de serem resolvidos com relativamente pouco tempo ou dinheiro; e "injustamente negligenciados", sofrendo de um subinvestimento grave e desproporcional. Embora o movimento incentive as pessoas a aplicarem essa estrutura para determinar os próprios problemas, ele também oferece recomendações sobre quais seriam, em sua visão, os mais relevantes. "Eu e outros membros da comunidade[6] do altruísmo eficaz convergimos para três questões morais que consideramos excepcionalmente importantes e que obtêm uma pontuação extraordinariamente alta nessa estrutura", disse MacAskill em um TED Talk, em 2018.

O primeiro problema é melhorar a saúde global, por exemplo, distribuindo mosquiteiros baratos e eficazes para prevenir a malária. O segundo é abolir a pecuária industrial, o que poderia melhorar a vida de bilhões de animais "por apenas alguns centavos por animal". O terceiro envolve os riscos existenciais, aqueles que têm um valor esperado drasticamente negativo porque, não importa quão improváveis sejam, podem destruir toda a humanidade e interromper todo o valor futuro que seria gerado para o restante da civilização. Dentro dessa terceira categoria, o movimento aponta recomendações mais específicas sobre o que constitui um risco existencial: pandemias globais, guerra nuclear e inteligência artificial rebelde.

Com a identificação da teórica IA rebelde como um risco existencial, o AE passou a defender o mesmo tipo de segurança em IA que estava entrelaçada no DNA da OpenAI desde o início e que desempenhou um papel central no Divórcio. Amodei e os colegas que fundariam a Anthropic discordavam fundamentalmente de Altman e de outros executivos da OpenAI quanto à seriedade com que se deveria tratar a possibilidade de a IA devastar a civilização. Para Amodei, que levava esse risco a sério, os comportamentos de Altman — a falta de transparência no acordo com a Microsoft; a aparente compulsão de dizer às pessoas o que elas queriam ouvir para obter a concordância delas, apenas para que só depois percebessem as mudanças de rumo — não eram só as maquinações típicas de um executivo do Vale do Silício, mas sinais alarmantes e imorais que poderiam comprometer o destino da humanidade. À medida que a Anthropic se consolidava, ela se amparava nesta diferença quanto à reputação: enquanto a OpenAI de Altman brincava de forma imprudente com o futuro da humanidade, a Anthropic se apresentava como a empresa de princípios, aquela que priorizava a segurança

da IA.

Em 2021, quando os irmãos Amodei anunciaram a Anthropic, o interesse nessa ideologia catastrófica e existencialista sobre a segurança da IA já vinha crescendo, impulsionado principalmente pela rápida expansão da esfera de influência do AE. O que era uma filosofia de nicho havia se transformado em um movimento mainstream, graças ao influxo de recursos de bilionários do setor de tecnologia.

Uma década antes, Dustin Moskovitz, cofundador do Facebook, e sua esposa, a ex-jornalista Cari Tuna,[7] haviam criado a organização sem fins lucrativos Good Ventures, com o objetivo de doar a maior parte de sua fortuna. Naquele período, Holden Karnofsky, futuro marido de Daniela Amodei, dirigia outra entidade, a GiveWell, fundada por ele em 2007 após deixar o fundo de hedge Bridgewater Associates. Movidas pelo mesmo desejo de distribuir dinheiro com métodos baseados em evidências, a Good Ventures e a GiveWell formaram, em 2011, uma parceria que mais tarde receberia o nome de Open Philanthropy. A partir daí, passaram a ampliar o financiamento para as principais áreas temáticas recomendadas por MacAskill — em especial, subsídios para pesquisas em segurança da IA, orientados pela estrutura conceitual do AE. A Open Philanthropy tornou-se uma organização independente em junho de 2017.[8]

Mais recentemente, um novo bilionário da tecnologia entrou em cena: Samuel Bankman-Fried, uma estrela em rápida ascensão graças ao grande sucesso como cofundador da corretora de criptomoedas FTX e da empresa de negociação Alameda Research. Bankman-Fried, ou SBF,[9] como é amplamente conhecido, atribuía ao AE a própria história. Formado em física pelo MIT, ele afirmava que pretendia seguir carreira acadêmica antes de ser convencido por MacAskill, durante um almoço, da superioridade moral de "ganhar para doar". A partir daí, SBF decidiu enriquecer o máximo possível para, finalmente, como prometia, investir toda a fortuna em filantropia.

À medida que acumulava sua fortuna[10] em um período notavelmente curto, SBF passou a doar dezenas de milhões de dólares para candidatos políticos, tanto democratas quanto republicanos, entre eles, o primeiro candidato apoiado pelo movimento do AE em 2022, no sexto distrito congressional do Oregon (que acabou não vencendo as primárias). Paralelamente, a bolsa de valores de SBF firmou acordos generosos que somavam bilhões em marketing esportivo, envolvendo atletas de ponta, como Tom Brady e Stephen Curry, e modalidades de alto prestígio, como a Fórmula 1. No movimento do AE, ele injetou não apenas dinheiro, mas também o brilho de sua crescente notoriedade. Quanto mais rico e famoso se tornava, maior era a visibilidade conferida à ideologia e ao seu cofundador, MacAskill. No início de 2022,[11] SBF anunciou a criação do próprio projeto filantrópico alinhado ao AE, o FTX Future Fund, com o objetivo de distribuir pelo menos 100 milhões de dólares, e até um bilhão, até o final daquele ano.

Em grande parte graças à Open Phil e ao FTX Future Fund, os anos de 2021 e 2022 registraram um salto significativo no fluxo de caixa destinado às pesquisas de segurança de IA apoiadas pelo AE. De acordo com estimativas compiladas por um membro da comunidade de AE e por dados da Open Phil, [12] o financiamento ultrapassou 100 milhões de dólares em cada um dos dois anos, depois de ter registrado, em média, menos da metade desse valor nos sete anos anteriores. Além de tê-la reforçado, esse influxo foi impulsionado pela crença progressiva de que o salto dramático nas capacidades do GPT-2 para o GPT-3 tornara a prevenção de riscos teóricos de uma IA descontrolada, e até riscos existenciais, mais urgente do que nunca. Cada vez mais pessoas passaram a aderir a esses projetos de segurança da IA, atraídas tanto pelos incentivos financeiros quanto pela ideologia, enquanto o número de integrantes do movimento do AE mais amplo crescia rapidamente. O AE sempre defendera a importância da preparação para pandemias e, em meio a uma pandemia real, sua aparente presciência conquistava novos adeptos. Ao mesmo tempo, o impacto psicológico de uma catástrofe global deixava muitas pessoas ansiosas e desorientadas, em busca de propósito, e elas encontraram no AE um caminho possível.

O crescente número de membros da comunidade de segurança da IA, que une a vertente apoiada pelo AE a outras correntes de pensamento catastrófico, existencial e focado no risco, ampliou as fileiras da Anthropic e reabasteceu o clã de Segurança da OpenAI. Os fóruns on-line do AE e da segurança da IA, [13] principais espaços de propagação, troca e debate entre esses movimentos sobrepostos, incentivavam os adeptos a trabalharem nos grandes laboratórios de IA, especialmente naqueles que acreditavam necessitar de maior vigilância em segurança, como OpenAI e DeepMind, com o objetivo de moldar e influenciar suas trajetórias. A entrada de novos integrantes na segurança da IA também popularizou o léxico da comunidade dentro da indústria de tecnologia como um todo. A velocidade com que alguém acredita que a IA avançará e atingirá marcos importantes, como a IAG, constitui sua "AI timeline" (ou "linha do tempo da IA"). A probabilidade de que a IAG provoque resultados catastróficos, como a morte da *maior parte* da população ou resultados existenciais, ou seja, a completa e total extinção da humanidade, é chamada de *p(doom)*, abreviação de "probabilidade de destruição", no original em inglês.[14] *Hardware overhang* (ou "excesso de hardware"), termo presente no roteiro de pesquisa da OpenAI para 2021, é outra entrada do dicionário, assim como *AI takeoff* (ou "decolagem da IA"), o processo pelo qual a IAG se aprimoraria até alcançar a superinteligência e, assim, ultrapassar a humanidade. Já o *acceleration risk* (ou "risco de aceleração") refere-se ao perigo de desencadear uma competição acirrada entre empresas ou países, levando a uma corrida potencialmente perigosa no avanço da IA e a um encurtamento do cronograma de desenvolvimento.

No entanto, para um movimento que professava o pensamento independente, o AE avançava rapidamente na direção oposta. Pessoas atraídas

pela premissa inicial eram logo doutrinadas seguindo um conjunto mais amplo de dogmas, alimentadas pela promessa de mais oportunidades e recursos, além de uma rede de contatos restrita que borrava os limites entre o pessoal e o profissional. No Vale do Silício, em especial, os integrantes do AE trabalhavam quase exclusivamente com outros integrantes do AE; viviam, festejavam, namoravam e dormiam, em grande parte, apenas entre si. Combinados ao sexismo profundamente enraizado na indústria de tecnologia e às subculturas poliamorosas tradicionais da Bay Area,[15] esses elementos criavam um fervor característico de uma espécie de culto que, em sua pior expressão, podia se transformar em um caldeirão tóxico de sexo, dinheiro e poder, cenário que vinha levando o AE a enfrentar um número crescente de acusações de assédio e abuso sexual.

Em novembro de 2022, a queda espetacular de SBF — com o colapso da FTX, uma longa série de condenações por fraude e uma sentença de 25 anos de prisão — seria, para muitos, um sintoma da podridão que havia se espalhado pelo movimento. Tão rapidamente quanto ganhou popularidade, o AE perdeu prestígio na indústria de tecnologia, e inúmeras pessoas passaram a se distanciar do movimento quase de imediato.

Contudo, mesmo sem o rótulo, persistiram as redes sociais do movimento, os valores, os jargões e a proeminência que o AE conferiu às questões existenciais de segurança da IA. Dessa permanência também emergiria uma força contrária: o e/acc (pronunciado "ee-ack"), ou aceleracionismo eficaz. O que começou, em grande parte, como uma sátira ao movimento AE rapidamente demonstrou um espírito oposto: enquanto o AE e a comunidade mais ampla de segurança da IA defendiam posições extremas voltadas a desacelerar, e até frear, o desenvolvimento da IA, ou, na visão de Amodei, acelerar a pesquisa enquanto restringiam sua adoção, o e/acc abraçaria uma visão maximalista de pisar fundo no acelerador em ambos os casos. Para os adeptos, não apenas o progresso tecnológico é universalmente benéfico, como é um imperativo moral avançar o mais rápido possível. Assim, os dois grupos passaram a ser conhecidos coloquialmente como pessimistas (*doomers*) e aceleracionistas (*boomers*).

Dentro dessa bolha, alguns começaram a enxergar a Anthropic e a OpenAI como os rostos respectivos de cada movimento. Outros passaram a ver a OpenAI como um campo de batalha entre ideologias polarizadas, uma organização originalmente enraizada no pensamento pessimista de uma instituição sem fins lucrativos, mas que estaria sendo capturada pelos aceleracionistas, com sua ênfase crescente, por meio do braço com fins lucrativos, em ganhar dinheiro. Havia incerteza quanto à lealdade de Altman; em diferentes momentos, ele parecia simpático a ambos os lados. Os mais benevolentes o viam como alguém tentando ocupar um meio-termo, encarregado da difícil tarefa de representar todas as perspectivas dentro de sua empresa. No entanto, após o Divórcio e a briga pessoal entre Altman e os cofundadores da Anthropic, cada vez mais pessimistas passaram a vê-lo da

pior maneira possível. Muitos dos elementos que colocaram a OpenAI no mapa, e impulsionaram o crescente sucesso comercial, haviam começado como projetos de segurança de IA: leis de escala, geração de código, aprendizagem por reforço a partir do feedback humano e a combinação desses três fatores em modelos linguísticos incrivelmente convincentes, seguidos por modelos multimodais. Para muitos pessimistas, o trabalho estava sendo cooptado e distorcido para fins completamente antitéticos em relação a seus valores fundamentais. Na visão deles, era Altman quem promovia essa apropriação e distorção e, por isso, ele tinha se tornado um mentiroso patológico, um manipulador abusivo e uma ameaça à humanidade.

Em pouco tempo, o conflito entre essas ideologias polarizadas dentro da OpenAI, e no ambiente que a cercava, ameaçaria destruir a própria empresa que mais havia contribuído para definir o tom da nova era do desenvolvimento da IA. No entanto, embora cada ideologia professasse ser a antítese da outra, ambas, na verdade, pregavam a partir da mesma bíblia.

As duas tratavam a IAG como uma conclusão cada vez mais inevitável, discutida com uma ferocidade quase religiosa; ambas se fixavam no longo prazo e reivindicavam uma autoridade moral para manter o desenvolvimento da IA sob o controle dos próprios adeptos. Enquanto uma alertava sobre fogo e enxofre, a outra seduzia com visões do paraíso.

No início de 2022, a OpenAI estava pronta para testar uma estratégia diferente de lançamento de produto, desta vez, com seu modelo texto-para-imagem. Ela não o esconderia por trás de uma API, nem entregaria o produto e a marca à Microsoft. A OpenAI faria o lançamento por conta própria e colocaria a tecnologia diretamente nas mãos dos consumidores. O modelo tinha até um nome atraente, criado pelos pesquisadores originais que o desenvolveram na empresa: DALL-E 2, uma referência ao artista surrealista espanhol Salvador Dalí e ao robô que dá nome ao filme da Disney/Pixar *WALL-E*.

O DALL-E surgiu de uma tendência no campo mais amplo da pesquisa em IA de desenvolver modelos multimodais — modelos que combinam ao menos duas "modalidades" distintas, como texto, imagens, som ou vídeo. Durante anos, o campo trabalhou para integrar as duas primeiras — linguagem e visão — de modo que um único modelo fosse capaz de relacionar palavras a informações visuais. Esse esforço foi impulsionado, em parte, pela disponibilidade de dados — texto e imagens são abundantes on-line e relativamente fáceis de processar —, mas também partia de uma hipótese científica: se a linguagem pura não basta para produzir inteligência em nível humano, a visão é provavelmente o segundo ingrediente mais poderoso.

Na OpenAI, inspirando-se nessa área, a equipe de Pesquisa seguiu a mesma progressão: depois dos modelos de linguagem, avançaram para modelos de texto e imagem e, de forma fundamental, concentraram-se em continuar usando Transformers para preservar a escalabilidade dos modelos.

Embora o primeiro Transformer tenha sido inicialmente projetado para funcionar melhor com texto, o Google introduziu, em 2020, o Vision Transformer, adaptando essa arquitetura ao processamento de imagens.

Em janeiro de 2021, a OpenAI apresentou dois novos modelos baseados em Transformers. O primeiro, chamado CLIP,[16] desenvolvido mais uma vez por Alec Radford, utilizava o Transformer original combinado ao Vision Transformer para gerar legendas detalhadas para imagens. O segundo, DALL-E 1,[17] criado por Aditya Ramesh, pesquisador formado pela Universidade de Nova York e que trabalhou por um período com Yann LeCun, da Meta, treinava um Transformer de 12 bilhões de parâmetros para receber texto e gerar imagens inéditas.

Em uma postagem no blog, a OpenAI destacou as capacidades do DALL-E 1 com uma série de prompts divertidos, incluindo "uma poltrona de abacate", que gerou diversas poltronas verdes e marrons inspiradas esteticamente em abacates. As imagens eram ligeiramente desfocadas e caricaturais, um artefato do processo de treinamento usado por Ramesh para produzir o modelo. Ele havia compactado 250 milhões de imagens para alimentar o Transformer, perdendo parte dos detalhes de alta resolução.

Quando a equipe começou a trabalhar no DALL-E 2, um novo método para gerar imagens estava ganhando tração. Conhecida como difusão, essa técnica inspirada na física permitia que Transformers aprendessem com mais precisão as correlações entre pixels em uma vasta gama de imagens. A ideia original veio de um artigo de 2015 escrito por pesquisadores de Stanford e Berkeley.[18] Cinco anos depois, Jonathan Ho, então estudante de pós-graduação em Berkeley orientado por Pieter Abbeel, um dos primeiros pesquisadores da OpenAI, popularizou a técnica ao reformulá-la de maneira engenhosa, produzindo imagens com fidelidade muito mais alta.[19] Ho também demonstrou que os modelos de difusão reconheciam imagens melhor do que os sistemas de visão computacional existentes. As descobertas podiam ser comparadas aos resultados de Radford com o GPT-1: ao aprender a sintetizar imagens convincentes — o equivalente a gerar frases semelhantes às criadas por humanos —, os modelos de difusão acabavam capturando padrões em seus dados de treinamento de forma tão profunda que se tornavam capazes de realizar uma gama mais ampla de tarefas de processamento visual.

A OpenAI mudou de estratégia para construir o DALL-E 2, combinando difusão com o CLIP de Radford.[20] Ramesh e outros pesquisadores[21] foram ampliando gradualmente o modelo e adicionaram a capacidade de preenchimento, permitindo que um usuário apagasse o cabelo de uma pessoa em uma foto e mudasse a cor, ou selecionasse um prado gramado em uma imagem e o preenchesse com zebras vagando. O uso da difusão gerou imagens muito mais nítidas e fotorrealistas; além disso, o método reduziu significativamente a quantidade de computação necessária para alcançar o mesmo desempenho do DALL-E 1.

Pesquisadores de fora da OpenAI reduziriam ainda mais a intensidade

computacional dos modelos de difusão.[22] O Stable Diffusion, popular gerador de imagens de código aberto, exigia apenas 256 Nvidia A100s[23] para ser treinado, graças a uma técnica revisada conhecida como difusão latente. Björn Ommer, professor da Ludwig Maximilian University, de Munique, cujo laboratório criou o Stable Diffusion, afirma ter desenvolvido a técnica após observar que os geradores de imagens estavam seguindo o mesmo caminho dos grandes modelos de linguagem, tornando-se obscenamente caros. "Estávamos presos em um trem que avançava rumo não apenas ao treinamento, mas também à conclusão, que, na prática, exigiria supercomputadores para funcionar; e milhões de dólares em investimentos", diz ele. "Nós nos perguntávamos se seria possível trazer a comunidade de pesquisa mais ampla de volta ao jogo e garantir que o campo da IA generativa não avançasse para um cenário em que apenas poucas grandes empresas de tecnologia teriam os recursos necessários para executar e hospedar esses modelos."

A OpenAI só adotaria a difusão latente muito mais tarde, o que deixou o DALL-E 2 e o DALL-E 3 muito mais caros em termos computacionais do que o Stable Diffusion ou o Midjourney, modelos que muitos usuários consideravam de qualidade superior. Esse foi apenas um exemplo de como, mesmo dentro do restrito campo da IA generativa, a escala não era o único caminho, nem necessariamente o mais eficiente, para alcançar capacidades mais avançadas de IA.

Com o notável salto de desempenho do DALL-E 2, a área de Produtos começou, no fim de 2021 e início de 2022, a trabalhar em diferentes ideias para sua comercialização.[24] A escolha recaiu sobre um aplicativo web chamado Labs, que permitiria aos usuários experimentar o modelo, e outros modelos futuros, diretamente pelo navegador. Tanto o chefe de produto, Fraser Kelton, quanto o vice-presidente, Bob McGrew, acreditavam que essa experiência interativa atenderia à demanda evidente revelada pelo GitHub Copilot de que as pessoas queriam se envolver diretamente com modelos de IA generativa. A solução também ajudaria a cumprir a missão da empresa — o DALL-E 2 era divertido e agradável, uma forma eficaz de suavizar o receio das pessoas em relação a sistemas de IA poderosos e pavimentar o caminho para que a OpenAI oferecesse mais benefícios de sua tecnologia em futuros lançamentos.

Com uma equipe de produtos ainda relativamente pequena, a empresa recrutou algumas pessoas para ajudar no design e no desenvolvimento do site. Para esses novos membros, vindos de ambientes corporativos mais tradicionais, a OpenAI parecia mais um laboratório de pesquisa universitário do que uma empresa. Os dias eram frequentemente dedicados à leitura de artigos acadêmicos e a debates teóricos, e não a revisar mock-ups de interfaces. Contudo, para alguns pesquisadores, a presença crescente da equipe de Produtos em suas reuniões de pesquisa produzia o efeito oposto. Já não eram

dias dedicados à investigação exploratória, com a discussão de ideias fundamentalmente novas sobre como criar um modelo multimodal melhor; naquele momento, uma parcela crescente do trabalho estava a serviço da comercialização, com o objetivo de descobrir maneiras de otimizar modelos existentes para atender às necessidades dos usuários.

Após a experiência com o combate ao abuso sexual infantil nos textos gerados no AI Dungeon, uma preocupação particular era a possibilidade de o DALL-E 2 ser usado para manipular material real ou criar material de abuso sexual infantil, ou CSAM, na sigla em inglês. Assim como ocorria com cada novo modelo GPT, os dados de treinamento de cada versão subsequente do DALL-E estavam se tornando mais poluídos. Para o DALL-E 2, a equipe de pesquisa firmou um contrato de licenciamento com a plataforma de fotos Shutterstock e realizou uma coleta massiva de dados no Twitter para complementar sua coleção existente de 250 milhões de imagens. O conjunto de dados proveniente do Twitter, em especial, estava repleto de conteúdo pornográfico. Diversos funcionários dedicaram esforços significativos para verificar e remover qualquer indício de material de abuso sexual infantil.

No entanto, após algumas discussões, os funcionários decidiram manter outros tipos de imagens sexuais, pois consideravam esse conteúdo parte da experiência humana. Por outro lado, manter essas fotos nos dados de treinamento significava que o modelo ainda seria capaz de produzir material de abuso sexual infantil. Da mesma forma que o DALL-E conseguia gerar uma poltrona de abacate tendo visto apenas abacates e poltronas, o DALL-E 2 e o DALL-E 3 podiam fazer o mesmo com crianças e pornografia para criar pornografia infantil, uma capacidade conhecida como "geração composicional".

Sem a filtragem dos dados para resolver a raiz do problema, o ônus passou a ser a criação de mecanismos de prevenção de abuso no modelo. Isso incluía filtros de moderação de conteúdo atualizados, que envolviam o próprio modelo para bloquear imagens de abuso, além de texto, bem como uma plataforma de monitoramento do comportamento dos usuários e uma chamada infraestrutura de banimento, composta de sistemas que suspendiam automaticamente as contas que atingissem certo limite de infrações repetidas. A empresa também contratou um novo chefe de Trust and Safety, Dave Willner, que, como um dos primeiros funcionários do Facebook, havia escrito os primeiros padrões de conteúdo da plataforma.

Mais tarde, durante o desenvolvimento do DALL-E 3, quando a necessidade de dados se tornou ainda maior, a equipe de pesquisa concluiu que imagens sexuais já não eram apenas um "extra", mas uma "necessidade". A proporção de conteúdo pornográfico na internet era tão alta que removê-lo reduzia o conjunto de dados de treinamento a ponto de degradar significativamente o desempenho do modelo. Em particular, essa remoção piorava a capacidade do modelo de gerar rostos de mulheres e de pessoas não brancas, devido à mesma descoberta feita por Deborah Raji quando era

estagiária na Clarifai: uma parcela significativa do conteúdo on-line que retrata esses dois grupos é sexualmente explícita. Pelas mesmas razões, os pesquisadores também deixaram no conjunto de dados alguns outros tipos de imagens perturbadoras.

Em dezembro de 2023, um engenheiro de IA desesperado da Microsoft, Shane Jones, descobriria as consequências dessas decisões.[25] Enquanto testava o Copilot Designer, o gerador de imagens da Microsoft baseado no DALL-E 3, ele ficou horrorizado ao notar a rapidez com que a ferramenta exibiu imagens ofensivas e sexualizadas após pouquíssimos prompts.

Ao adicionar o termo "pró-escolha" ao prompt, Jones descobriu que o sistema produzia cenas de um demônio devorando um bebê e o que parecia ser uma broca com a inscrição "pró-escolha" sendo usada para mutilar um bebê. Quando solicitou apenas "um acidente de carro", e nada mais, a ferramenta gerou imagens de mulheres sexualizadas ao lado de acidentes violentos, incluindo uma de lingerie ajoelhada ao lado de um veículo destruído, como a CNBC posteriormente constatou nos próprios testes. Durante três meses, Jones pediu aos executivos da Microsoft que retirassem a ferramenta do ar até que ela tivesse melhores proteções ou, ao menos, que alterassem sua classificação na loja de aplicativos do Android de "Livre" para uma classificação voltada para adultos. Depois que a Microsoft recusou a recomendação e a OpenAI não respondeu, ele enviou uma carta à Comissão Federal de Comércio dos Estados Unidos (Federal Trade Commission). "Eles não implementaram essas mudanças e continuam a comercializar o produto como 'Qualquer pessoa. Qualquer lugar. Qualquer dispositivo'", escreveu ele à Comissão. Esse problema, acrescentou, "era conhecido pela Microsoft e pela OpenAI antes do lançamento público do modelo de IA em outubro passado". A Microsoft não comentou sobre a questão ou o desfecho mais recente da carta de Jones.

À medida que o lançamento do DALL-E 2 se aproximava, a disputa entre a divisão de Produtos da OpenAI e o recém-reabastecido clã de Segurança ressurgiu.

Para os membros da área de Segurança, dispersos por várias equipes dentro da divisão de Pesquisa, o realismo sem precedentes do DALL-E 2 trazia uma vasta gama de incógnitas. Como ele poderia ser usado para produzir material de abuso sexual infantil ou deepfakes políticos? Para manipular e persuadir pessoas? Para abusar e prejudicar indivíduos ou para criar impactos sociais amplos de maneiras que estavam além da previsão e da imaginação da OpenAI? Eles pressionaram a empresa a não lançar o modelo sem testes mais rigorosos e evidências sólidas de que não causaria danos.

Para o pessoal de Produtos, a lista sempre crescente de preocupações mais uma vez parecia histérica e o nível de exigência para o lançamento, completamente irreal. Nenhum sistema poderia jamais garantir *zero* dano, especialmente um que permanecesse trancado em laboratório, sem qualquer

contato com usuários reais. Assim como a equipe de Segurança temia as limitações de previsão da OpenAI, a de Produtos acreditava que era exatamente por isso que o DALL-E 2 precisava ser lançado. Colocar modelos de IA em circulação de forma controlada, para obter feedback do mundo real, eliminaria parte desse trabalho de adivinhação e, portanto, era um passo necessário para melhorar a segurança.

No centro do conflito estava uma discordância cada vez mais intensa sobre o que, exatamente, a OpenAI era. Para o clã da Segurança, a OpenAI ainda era um laboratório de pesquisa idealista e sem fins lucrativos, com a obrigação primordial de, conforme estabelecido no estatuto, colocar o benefício da humanidade acima de quaisquer interesses comerciais. Sob essa premissa, os benefícios de reter os modelos pelo tempo necessário, para refletir sobre todas as possíveis desvantagens e pesquisar maneiras de mitigá-las, superavam em muito os custos. Para a divisão de Produtos, entretanto, a OpenAI precisava tomar decisões mais práticas, baseadas na maneira como o mundo funcionava. Permanecer líder em pesquisa de IA era essencial para a missão da empresa e para definir normas sobre o desenvolvimento da tecnologia. Isso significava tolerar certo grau de risco para agir de modo rápido, em especial, em meio a rumores de que o Google estava finalizando o próprio gerador de imagens, além de garantir a extraordinária quantia de capital necessária para continuar realizando pesquisas de ponta. Este último ponto exigia levantar fundos de investidores, o que, por sua vez, exigia trabalhar de boa-fé para desenvolver uma estratégia comercial que um dia oferecesse retorno a esses investidores.

As pessoas da área de Segurança eram "completamente ingênuas" sobre como as empresas e o mundo funcionam, comenta um ex-funcionário de Produtos.

"Bem, os riscos da missão de IAG proposta pela OpenAI são altos", afirma um integrante da Segurança. "Uma 'empresa normal' talvez não seja boa o suficiente."

Diferentes equipes estavam codificando esse conflito crescente nas métricas que usavam para avaliar seu desempenho. A divisão de Produtos, em uma operação de entrada no mercado em desenvolvimento, trabalhava para definir metas de crescimento de usuários e de receita. Já na divisão de Pesquisa, as várias equipes de segurança de IA se esforçavam para encontrar maneiras quantificáveis de medir o avanço, algo difícil de especificar por natureza. A segurança de IA ainda era uma disciplina relativamente jovem, sem referências claras ou estabelecidas. Em reuniões e no Slack, o pessoal da Segurança repetidamente manifestava à alta liderança preocupações sobre como esse desequilíbrio criava incentivos desalinhados: ter metas nítidas de crescimento e receita, sem algum tipo de contrapeso forte e comparável, estava levando a OpenAI a operar cada vez mais como uma organização que "avançava a qualquer custo sem medir consequências".

Em conversas privadas com a equipe de Segurança, Altman expressava simpatia pela perspectiva, concordando que a empresa não estava no caminho

certo com sua pesquisa de segurança de IA e que precisava investir mais nessa área. Em conversas privadas com a divisão de Produtos, porém, ele os incentivava a seguir em frente. Durante as reuniões do conselho, Altman concordou com Brockman quando este expressou frustração com a forma como algumas pessoas estavam usando a segurança da IA enquanto uma alavanca política para impedir o progresso em prol dos próprios interesses.

Cada vez mais, Mira Murati foi assumindo o papel de negociadora, suavizando as divergências entre as diferentes facções e buscando maneiras de conciliar as posições. No caso do DALL-E 2, ela chegou a um acordo: o aplicativo web seria lançado não como um produto, mas como uma "versão de teste discreta". Essa designação daria à OpenAI mais liberdade para impor restrições severas ao modelo, satisfazendo a equipe de Segurança, ao passo que permitiria testar um relacionamento direto com o consumidor e coletar feedback dos usuários, agradando a divisão de Produtos. A decisão também era prática. A OpenAI ainda não tinha a infraestrutura necessária para moderar o conteúdo das imagens geradas. Chamar o modelo de "versão de teste" e não cobrar por ele permitia à empresa usar bloqueadores amplos e diretos sem se preocupar em frustrar usuários pagantes, ganhando tempo para desenvolver filtros mais sofisticados. A empresa avançou com a implementação de uma série de mecanismos agressivos para a prevenção de conteúdos envolvendo abuso, incluindo a desativação da capacidade do DALL-E 2 de gerar rostos fotorrealistas ou editar fotos reais com rostos, uma forma de contornar completamente os problemas do material de abuso sexual infantil e da desinformação política.

Em março de 2022, a OpenAI lançou o DALL-E 2 pelo aplicativo web Labs, para delírio do público. Enquanto as pessoas se encantavam e se intrigavam com as capacidades do modelo, em um nível que superou as expectativas de muitos funcionários, o aplicativo logo se tornou viral nas redes sociais, produzindo uma profusão de arte gerada por IA louca, excêntrica e surreal. Foi um momento GPT-3, só que melhor. Em vez de interagir apenas com um pequeno grupo de desenvolvedores técnicos, a empresa alcançava uma base muito mais ampla e global de consumidores. Em tempo real, também podia responder ao feedback dos usuários com mudanças instantâneas no aplicativo Labs. "Isso é inebriante", Fraser Kelton recordaria mais tarde, em um podcast.[26]

Nos meses seguintes, a divisão de Produtos, que ainda não havia pensado muito sobre como monetizar o DALL-E 2, correu para transformar o aplicativo web em um produto pago. Ela passou a trabalhar com artistas e profissionais criativos de todo o mundo para incorporar o DALL-E 2 em suas práticas e lançou um programa beta que convidava um milhão de pessoas a acessarem o modelo, oferecendo créditos gratuitos para geração de imagens. Contudo, quando a OpenAI começou a cobrar, não foi o Google que se mostrou o principal desafio, embora a gigante realmente tenha seguido com seu modelo Imagen. Em vez disso, os maiores concorrentes foram dois

modelos de duas startups: Midjourney e Stable Diffusion, da Stability AI. Ambos eram gratuitos e tão bons, se não melhores, do que o DALL-E 2, além de terem menos medidas de segurança, permitindo que os usuários gerassem e editassem até mesmo rostos de políticos.

À medida que o DALL-E 2 rapidamente perdia força no mercado, a experiência deixou a equipe de Produtos com uma sensação incômoda de que havia desperdiçado uma grande oportunidade comercial — em parte porque o aplicativo era excessivamente restritivo. A equipe já estava em processo de desativar seus bloqueadores diretos e substituí-los por proteções mais direcionadas. Impulsionados pelo desejo de superar os concorrentes, os executivos pressionavam a equipe a desabilitá-los o mais depressa possível.

Para suspender o bloqueio da geração de imagens de rostos, a OpenAI desenvolveu um novo processo para prevenir e reprimir a geração de imagens prejudiciais de pessoas, incluindo material de abuso sexual infantil. Ela utilizou sistemas automatizados para detectar quando rostos estavam sendo gerados em contextos aceitáveis ou abusivos e, mais uma vez, contou com a ajuda de prestadores de serviços no exterior para moderar o conteúdo. Desta vez, eles trabalhavam da Índia, por meio de um fornecedor chamado Cogito, e revisavam não apenas grandes quantidades de texto, mas também imagens — artificiais e reais — do mesmo tipo de conteúdo sexual e violento que havia sido enviado aos funcionários da Sama. Ao analisarem o que poderia chegar a centenas de imagens por dia, os contratados enfrentavam dificuldade para distinguir entre conteúdo sexual envolvendo menores de dezessete anos e adultos de dezoito anos. Eles nem sempre conseguiam determinar se as imagens eram falsas ou reais.

O que, à primeira vista, parecia ser a meta mais simples do roteiro de pesquisa da OpenAI para 2021 acabou se revelando uma das mais difíceis: ampliar o GPT-3 em dez vezes usando o novo supercomputador da Microsoft, equipado com 18 mil Nvidia A100, no esforço de desenvolver o que se tornaria o GPT-4. Um terço da equipe responsável pelo escalonamento do GPT-3 havia deixado a empresa durante o Divórcio, levando consigo uma quantidade significativa de conhecimento técnico e institucional. Essencialmente, a OpenAI havia ficado sem dados.

Após o GPT-3, os pesquisadores passaram a buscar o máximo de dados possível, ampliando o acervo da empresa ao baixar todos os novos conjuntos disponíveis e coletar cada novo fórum on-line que encontravam e que não exibia avisos claros contra essa prática. No entanto, mesmo com as adições do vasto repositório do GitHub para o Codex e dos livros didáticos e manuais de programação, ainda não era suficiente.

Diante de uma batalha árdua, a situação tinha todas as características de um projeto com a cara de Greg Brockman. Não apenas canalizaria a atitude determinada e a genialidade em codificação, como também concentraria a energia, para o bem do restante da empresa, em algo produtivo.

Depois que Altman assumiu, dispensando Brockman de suas responsabilidades gerenciais, este acabou retornando ao papel de colaborador individual, sem subordinados. No entanto, como presidente titular e um dos cofundadores da OpenAI, manteve uma influência extraordinária sobre os funcionários e sobre a direção estratégica da empresa. À medida que a OpenAI se profissionalizava e implementava processos corporativos mais padronizados, afastando-se dos dias mais livres e improvisados de uma startup em estágio inicial, a combinação da baixa responsabilidade formal e da alta autoridade de Brockman começou a se tornar um problema.

Assim como nos dias de faculdade e na Stripe, Brockman não era muito fã de instituições ou processos, pois tinha uma energia inquieta e obsessiva. Raramente participava de reuniões e definia a própria agenda, muitas vezes preferindo programar por dezenas de horas seguidas, com poucas pausas para comer ou dormir. Com o projeto certo, os efeitos eram milagrosos: sua produtividade intensa impulsionava o progresso.

Por outro lado, quando ficava ocioso, ele tendia a deixar um rastro de destruição, surgindo em projetos diversos para interferir e atrapalhar planos de longa data com mudanças de última hora. Às vezes, quando os funcionários resistiam, ele recorria a apelos emocionais cada vez mais intensos à sua liderança para conseguir o que queria.

Brockman geralmente conseguia o que queria. Para grande frustração e perplexidade de outros executivos, Altman era permissivo com seu comportamento. Além disso, ele também conseguia influenciar Altman a interferir e atrapalhar coisas a seu favor, apenas para satisfazê-lo. Uma suposição bastante difundida sobre a razão disso era a seguinte: embora Altman fosse o chefe de Brockman como CEO, Brockman também tinha autoridade sobre Altman como integrante do conselho. Era uma estrutura estranha e confusa que, no fim das contas, não deixava nada nem ninguém a cargo de conter ou responsabilizar Brockman.

A alta liderança mudou a função, o escopo e a quem Brockman se reportava várias vezes, na tentativa de encontrar o melhor lugar para ele. Como em tantos outros casos, a responsabilidade acabou recaindo sobre Murati, que se tornou a gerente de Brockman. Quando ela tentava dar feedback, ele parecia receptivo, mas, nos pontos em que discordava, reclamava diretamente com Altman. Aos poucos, ela desistiu de tentar promover mudanças por meio de feedback e, em vez disso, passou a dedicar um tempo significativo tentando encontrar projetos nos quais Brockman pudesse ajudar, e não causar caos, e, juntamente com McGrew, passou a remendar as rupturas que ele provocava em várias partes da empresa.

Com diversos obstáculos ainda por serem superados no caminho do desenvolvimento do GPT-4, as estrelas se alinharam.

Para resolver o gargalo de dados da OpenAI, Brockman recorreu a uma nova fonte: o YouTube.[27] A OpenAI havia evitado essa opção anteriormente —

usar o YouTube para treinar os modelos da empresa, como o próprio CEO da plataforma confirmaria mais tarde, violava os termos de serviço do site.[28] Contudo, sob a nova pressão por mais dados, a pergunta passou a ser outra: o YouTube, ou sua controladora, o Google, realmente aplicaria esses termos? Se o Google tomasse medidas severas contra a OpenAI, isso poderia comprometer a capacidade da própria empresa de coletar dados de outros sites para treinar o próprio grande modelo de linguagem. Brockman estava disposto a correr o risco.

Com uma pequena equipe, ele começou a coletar vídeos do YouTube, acabando por compilar mais de um milhão de horas de filmagens, de acordo com o *New York Times*. Em seguida, usou uma ferramenta de reconhecimento de voz chamada Whisper, desenvolvida por Radford, para transcrever os vídeos para o GPT-4.[29]

Depois veio o treinamento. Para treinar o GPT-3, a equipe da Nest havia projetado uma plataforma de software personalizada. Como a maioria de seus criadores estava na Anthropic, eles já não estavam por perto para explicar como tudo funcionava. Por uma questão de orgulho, alguns líderes também não queriam depender do legado deixado pela equipe que havia migrado para a Anthropic. Brockman então desapareceu em seu buraco de codificação e desenvolveu uma nova plataforma. Depois, com vários outros pesquisadores, [30] incluindo Jakub Pachocki e Szymon Sidor, os cientistas poloneses de quem ele havia se aproximado durante o projeto Dota 2, Brockman supervisionou o treinamento do GPT-4. Só o pré-treinamento levou três meses.

No início, o GPT-4 parecia uma decepção. "Era um modelo descontrolado, que, em certo sentido, se comportava muito mal", diz um pesquisador. "Como a qualidade média dos dados era péssima e o modelo era bastante poderoso e sensível ao contexto, ele produzia respostas sem sentido." No entanto, Brockman seguiu em frente, reunindo recursos para melhorar o modelo com colaboradores humanos que realizavam aprendizagem por reforço a partir do feedback humano. A cada semana, os resultados iam se tornando melhores, até que o desempenho começou, de fato, a impressionar as pessoas dentro da empresa.

O GPT-4 passou a ter recursos multimodais integrados e, contrariando as avaliações internas da OpenAI, gerava um código mais refinado do que nunca e se mostrava muito mais ágil em reconhecer a intenção do usuário e entregar respostas úteis. Em uma demonstração impressionante dessas habilidades, Brockman transmitiu ao vivo um teste no qual pediu ao GPT-4 que transformasse em código a foto de um simples esboço de uma página da web desenhado em seu caderno. "Meu site de piadas", ele escreveu no topo do esboço. Abaixo, acrescentou: "[piada muito engraçada!]" e "[aperte para revelar a piada]". Em menos de meio minuto, o modelo converteu o rabisco em um código funcional, organizando a primeira linha como título, substituindo a segunda por uma piada e reconhecendo a terceira como um botão.

Contudo, quando a OpenAI começou a apresentar o modelo em círculos confiáveis, incluindo investidores e alguns clientes selecionados, pelo menos uma pessoa não se mostrou nem um pouco impressionada. Mais uma vez, era o sempre exigente Bill Gates.

Em junho de 2022, após receber uma demonstração do GPT-4, Gates expressou decepção com o que considerava um progresso insuficiente em relação ao GPT-2. Apesar de o modelo ser significativamente maior e mais fluente, ele ainda achava que se comportava como "um sábio idiota",[31] incapaz de lidar com problemas científicos complexos. Gates disse à equipe que só começaria a levar o modelo a sério quando o GPT-4 obtivesse uma nota 5 no exame de Biologia AP — AP Bio —, porque acreditava que aquela prova avaliava o pensamento científico crítico, e não a simples memorização de fatos. "Pensei: 'Ok, isso me dará três anos para trabalhar com o HIV e a malária'", contou ele mais tarde em seu podcast.[32]

Brockman interpretou a observação de Gates como um desafio. Ele imediatamente entrou em contato com Sal Khan, CEO da plataforma de educação on-line Khan Academy, e pediu para utilizar o grande repositório de questões de biologia avançada da empresa como dados de treinamento. Khan estava cético, mas concordou, com a condição de que sua plataforma tivesse acesso ao modelo. Brockman também reuniu uma equipe de funcionários para construir uma interface de usuário especial para a nova demonstração destinada a Gates.

No fim de agosto, para grande surpresa de Gates, Altman e Brockman voltaram a procurá-lo. No mês seguinte, durante um jantar na casa do fundador da Microsoft, com cerca de trinta convidados, os dois executivos da OpenAI, acompanhados de outros membros da empresa, apresentaram a Gates uma série de demonstrações altamente refinadas do GPT-4, projetadas para impressioná-lo. O ponto culminante foi o desempenho do modelo na AP Bio: ele acertou 59 das 60 questões de múltipla escolha e produziu respostas impressionantes para seis questões discursivas. Um especialista externo avaliou o resultado como uma nota 5 de 5. Gates não conseguia acreditar. O choque e os elogios relatados pelos presentes e imediatamente retransmitidos ao restante da empresa se espalharam como fogo pelos corredores, incitando um nível eletrizante de energia. Essa demonstração, disse Gates, foi uma das duas apresentações mais impressionantes que ele já havia visto na vida.[33]

Em reuniões com os funcionários, Altman continuou a alimentar o entusiasmo. "Startups que fazem coisas notáveis precisam de um milagre", disse ele. "Nós acabamos de ter nosso milagre." Muitos funcionários acreditaram nisso, impressionados com a importância do que haviam conquistado. O novo nível de desempenho do GPT-4 convenceu a liderança da OpenAI de que era hora de começar a trabalhar rumo a um dos objetivos de longo prazo de Altman: um assistente de IA como Samantha, personagem de *Her*, filme de Spike Jonze, de 2013.

Durante anos, *Her* foi uma referência que Altman e outros cofundadores da OpenAI frequentemente invocavam de exemplo de como a IAG poderia um dia ser: um único modelo multimodal cuja interface de produto fosse tão natural que se tornasse imperceptível, e simplesmente encantasse o usuário.

"Acho que é porque era uma assistente maravilhosamente integrada à vida", diz um ex-funcionário ao explicar por que o filme era uma referência tão importante. "O arco positivo da história, antes de tudo se desenrolar, é uma ótima metáfora para a evolução da IA na sociedade."

A equipe de pesquisa de John Schulman começou a reaplicar no GPT-4 o trabalho de chatbot com RLHF inspirado no InstructGPT, antes usado no GPT-3.5, para servir como o software central do produto que a liderança passou a chamar de Superassistant. Brockman e Fraser Kelton reuniram uma equipe heterogênea, com menos de dez pessoas vindas de diferentes áreas da empresa, para debater e prototipar ideias para a interface do produto.[34] Uma pessoa do time de supercomputação, normalmente responsável apenas pela infraestrutura, passou a trabalhar, nas noites e fins de semana, em um aplicativo para iOS que permitia conversar com o modelo. Outra sugeriu usar o Whisper para adicionar uma interface de voz, permitindo que usuários falassem com o sistema sem precisar digitar. Uma terceira, do time de inferência, propôs criar uma extensão do Chrome que ajudasse a resumir páginas da web com o Superassistant enquanto o usuário navegava. Uma quarta pessoa começou a desenvolver um bot de reunião, capaz de participar de videochamadas e enviar aos usuários um resumo do que havia acontecido.

À medida que o ritmo de trabalho aumentava, também crescia a empolgação com as possibilidades. Naquele verão, quando um grupo de pesquisadores de IA, incluindo Barret Zoph, Luke Metz e Liam Fedus, deixou o Google para fundar a própria startup de assistente digital, Altman os convenceu a desenvolver essa ideia dentro da OpenAI. Eles se juntaram à equipe de Schulman, instalando-se lado a lado no escritório com a equipe do Superassistant, para impulsionar a pesquisa e acelerar o desenvolvimento. Em um estado de fluxo intenso e eletrizante, pessoas das divisões de Produtos e de Pesquisa trabalhavam mais próximas do que nunca para lançar um novo produto.

Quando a OpenAI apresentou o GPT-4 à Microsoft, Satya Nadella, Kevin Scott e outros executivos da gigante tecnológica ficaram igualmente entusiasmados. O Codex havia provado que as tecnologias da OpenAI podiam ter apelo comercial, mas o GPT-4 representava algo muito maior. De modo geral, o modelo superava o desempenho de vários sistemas de IA que a Microsoft havia desenvolvido, e fazia muito mais, incluindo responder a perguntas com um alto grau de contextualização e clareza. Isso abriu uma gama de possibilidades para criar interfaces conversacionais, ou Copilots, para todos os produtos da Microsoft. A partir daquele momento, seria possível permitir que os usuários conversassem com o mecanismo de busca Bing, que enfrentava dificuldades, ou pedissem ao pacote Office, em linguagem natural,

para transformar um documento do Word em uma apresentação do PowerPoint. A Microsoft também poderia oferecer Copilots personalizados diretamente aos clientes de seu serviço de nuvem. Isso, sem dúvida, transformaria a empresa em uma líder em IA, finalmente capaz de competir de igual para igual com o Google.

Nos meses seguintes, Nadella desbloquearia o terceiro investimento da Microsoft na OpenAI e garantiria a continuidade do acesso exclusivo aos pesos dos modelos da empresa para integrá-los aos produtos da gigante tecnológica. Em janeiro de 2023, as empresas revelariam o valor do aporte: 10 bilhões de dólares.

Inicialmente, os executivos da OpenAI pretendiam lançar o GPT-4 no meio do segundo semestre de 2022. No entanto, o prazo acabou se revelando uma fantasia. Perto do fim agosto, a empresa estava longe de estar pronta para lançar um novo produto comercial em qualquer uma de suas frentes. A equipe de produto precisava de mais tempo para aperfeiçoar a interface; a equipe de infraestrutura ainda precisava distribuir o espaço de servidor; e o próprio modelo exigia mais trabalho para ajustar seu comportamento.

Naquele momento a OpenAI havia formado com a Microsoft um comitê chamado Conselho de Segurança de Implantação (DSB, em inglês), com três representantes de cada empresa, responsável por avaliar os modelos de ponta da OpenAI e determinar se estavam prontos para serem lançados. Os representantes da OpenAI eram Altman, Miles Brundage, chefe de pesquisa de políticas, e Jan Leike, líder de alinhamento, que supervisionava o desenvolvimento contínuo de técnicas de segurança de IA, como o RLHF. Tanto a pesquisa de políticas quanto o alinhamento haviam se tornado os novos bastiões emergentes do clã de Segurança, assim o DSB instituiu uma estrutura formal de governança para resolver os antigos debates entre Produtos e Segurança. Após uma análise preliminar, o DSB deu ao GPT-4, o primeiro avaliado por essa estrutura, uma aprovação condicional: o modelo poderia ser lançado, desde que passasse por testes extensivos e ajustes adicionais de segurança.

Os executivos concordaram com o novo prazo para o início de 2023. Até lá, queriam que o Superassistant também estivesse pronto. A OpenAI lançaria o GPT-4 na API lado a lado com o produto voltado para o consumidor, ambos equipados com o GPT-4. Para os funcionários, Altman apresentou a decisão de adiar o lançamento do modelo como prova da abordagem cautelosa e orientada para a segurança da OpenAI. A empresa estava levando tempo para implementar o sistema, afirmava ele. Na realidade, a liderança nem sequer tinha certeza se o lançaria.

Na divisão de Produtos, a maioria interpretou a fala de Altman como um sinal de que seria necessária uma circunstância realmente extraordinária para que a liderança interrompesse o lançamento, mas que, se tal situação surgisse, eles estariam dispostos a fazer isso. Para muitas pessoas dentro do clã de

Segurança, porém, as palavras de Altman tiveram outro impacto: a suposição era de que somente depois que o GPT-4 passasse por todas as verificações a liderança daria luz verde para o lançamento.

Em reuniões privadas, Altman reforçava as percepções de cada lado. Para ambos, ele levantava a preocupação de que o GPT-4 pudesse "acordar o Google", evocando um cenário em que, no dia do lançamento, caso o modelo chamasse a atenção de Larry Page e Sergey Brin, os dois finalmente superassem a burocracia política interna de sua empresa. Os executivos poderiam cancelar todas as reuniões, fazer um retiro de um dia, consolidar talentos e recursos e usar todo o poder computacional da empresa para recuperar o atraso.

Contudo, enquanto para a divisão de Produtos esse risco era motivo para continuar com a intensidade dos preparativos, mantendo as capacidades do GPT-4 em segredo pelo maior tempo possível, Altman usava, com o clã de Segurança, o mesmo argumento para enfatizar sua cautela. "Minha principal preocupação com a segurança é o risco de aceleração", dizia ele, adotando o vocabulário deles. Tradução: despertar o Google poderia desencadear uma dinâmica de corrida para o fundo do poço.

Mesmo com o atraso, a divisão de Produtos se esforçava para cumprir o prazo de início em 2023. Em particular, a equipe de Trust and Safety de Dave Willner ainda era pequena e operava com uma infraestrutura subdesenvolvida. Se a OpenAI planejava lançar o produto Superassistant diretamente ao consumidor, construído sobre um modelo muito mais poderoso, ela precisava de uma operação de prevenção de abusos e fiscalização muito mais madura.

Willner correu para contratar vários colaboradores experientes em Trust and Safety vindos do Google e da Meta. Ele encarregou um deles de investigar os riscos desconhecidos — abusos dos quais a equipe ainda não tinha conhecimento e que precisariam ser descobertos por meio de monitoramento cuidadoso e análise de dados. Encarregou outro dos riscos conhecidos, ou da chamada moderação em escala — abusos que a empresa já havia identificado como violações e que precisariam ser sinalizados, revisados e punidos, por exemplo, com a suspensão de contas de infratores reincidentes.

Enquanto corriam para definir as políticas de produto e a infraestrutura de monitoramento e fiscalização, os integrantes da equipe em formação tinham uma constante sensação de incerteza sobre como deveria funcionar uma área de Trust and Safety em uma empresa de IA, e em que aspectos isso diferia do trabalho em organizações de pesquisa ou de redes sociais. Eles também se perguntavam se os preparativos para o GPT-4 eram realmente adequados. A área de Trust and Safety tradicionalmente se concentrava na prevenção de uma série previsível de abusos na internet, como fraudes, crimes cibernéticos e interferência em eleições. No entanto, em meio à correria, surgia a dúvida de como seu trabalho se relacionava com os das equipes de segurança de IA, dentro do alinhamento liderado por Leike, e da pesquisa de políticas de

Brundage, que frequentemente discutiam cenários catastróficos e desconhecidos. A OpenAI rotulava todas essas equipes sob o mesmo guarda-chuva de "segurança", mas elas pareciam falar línguas diferentes. Embora o vocabulário referente ao risco existencial da IA, popularizado pelo AE, tivesse se tornado comum no campo, funcionários provenientes de empresas de tecnologia tradicionais nunca tinham ouvido falar de linhas do tempo da IA ou quantificado sua probabilidade de destruição. Mesmo palavras compartilhadas entre os grupos, como *risco* e *dano*, pareciam carregar conotações distintas.

Com a chegada de um número cada vez maior de pessoas sem formação específica em IA para apoiar as necessidades crescentes da divisão de Produtos, a capacidade de transpor essa barreira cultural — entre aqueles imersos na mentalidade pessimista-aceleracionista focada no risco existencial e aqueles que operavam sob uma estrutura padrão de empresa de tecnologia — tornou-se um tipo especial de moeda política. Entre os colaboradores de Willner na área de Trust and Safety, aquele que conseguiu aprender com mais sucesso a linguagem da segurança em IA acabou se tornando uma ponte entre esses mundos, trabalhando com diversas equipes a fim de virar o modelo "mais seguro" nos dois sentidos da palavra, usando RLHF para "alinhar" o modelo — no vocabulário da segurança de IA — e, assim, tornando-o melhor em recusar solicitações de usuários que violassem as políticas da plataforma da OpenAI, o trabalho típico de Trust and Safety. Entretanto, à medida que a "segurança" avançava em ambas as frentes, surgiu repentinamente uma nova diretriz dos executivos: suspender o processo de revisão de desenvolvedores que havia sido implementado pela primeira vez no lançamento da API do GPT-3.

Havia algum tempo, os executivos sentiam que a lista de espera tinha crescido desordenadamente e que o processo de revisão não estava acompanhando esse aumento. Os desenvolvedores reclamavam das longas demoras para obter acesso às tecnologias da OpenAI e enviavam e-mails para Altman, Brockman e Peter Welinder, ou os marcavam no X, para reclamar de atrasos em solicitações que, aos olhos dos executivos, eram totalmente inofensivas. Muitos na divisão de Produtos também acreditavam que a OpenAI estava restringindo o acesso aos benefícios de sua tecnologia de forma desnecessária, o que contradizia a missão da empresa.

Willner e sua equipe sempre resistiram. Se a OpenAI abandonasse o processo de inscrição e passasse a aprovar automaticamente os desenvolvedores, a empresa não teria alternativa real para moderar o uso de suas tecnologias. Caso migrasse para um sistema reativo, precisaria criar ferramentas robustas para sustentá-lo. Com o lançamento do GPT-4 se aproximando, porém, os executivos ignoraram as objeções: a OpenAI estava eliminando a revisão de desenvolvedores, e caberia à equipe de Trust and Safety descobrir a alternativa.

A equipe de Willner se apressou em elaborar uma proposta. A ideia era

deslocar uma parte maior da aplicação das políticas da empresa para o início do processo, confiando mais fortemente no RLHF para alinhar o GPT-4 e os modelos futuros. Todo o restante seria capturado e tratado por meio de uma fiscalização reativa: usando diferentes sinais de dados, como informações sobre o comportamento do aplicativo, os padrões de picos de tráfego e o número de vezes que os filtros de moderação de conteúdo foram acionados, para suspender automaticamente os infratores mais óbvios, enquanto encaminhava os casos limítrofes para moderadores humanos realizarem uma revisão manual.

Willner pediu aos executivos os recursos necessários para desenvolver adequadamente as ferramentas que permitiriam que o plano funcionasse. A OpenAI não tinha a maioria desses dados à disposição; a plataforma de monitoramento registrava apenas informações básicas sobre o tráfego que cada aplicativo enviava pelos servidores da empresa. Em alguns casos, ela nem sequer sabia o nome do desenvolvedor ou a finalidade do aplicativo.

Além de preencher essas lacunas, a equipe também precisava que os desenvolvedores atribuíssem um identificador exclusivo a cada um dos usuários. Isso era crucial para desagregar o tráfego de um aplicativo por usuários individuais e determinar se comportamentos infratores eram endêmicos na base de usuários ou cometidos apenas por alguns abusadores frequentes. Os executivos resistiram, preocupados com a possibilidade de que a implementação de um monitoramento tão minucioso gerasse atrito na experiência do desenvolvedor e, potencialmente, tornasse a OpenAI mais responsável por realizar o trabalho de Trust and Safety para cada aplicativo que utilizasse suas tecnologias, em vez de cada desenvolvedor assumir essa responsabilidade por conta própria.

No fim, a proposta de fiscalização reativa foi reduzida a algo muito mais limitado.

A visão limitada do comportamento dos usuários dos aplicativos deixou alguns integrantes da equipe de Trust and Safety preocupados. Um funcionário compartilhou suas inquietações com Brockman. O que ele mais temia, disse, era que as pessoas usassem o GPT-4 para gerar desinformação em grande escala e influenciar eleições.

O executivo tentou tranquilizá-lo. “Sim, isso é o que sempre dizemos e com o que sempre nos preocupamos”, respondeu. “Mas o que eu nunca vi é se isso está acontecendo de verdade.”

O GPT-4 não foi apenas um ponto de virada para Gates e para a Microsoft. Mais tarde naquele ano, após impressionar o bilionário filantropo, Altman e Brockman levaram Gates para uma apresentação ao conselho da OpenAI, na qual Brockman realizou uma demonstração ao vivo do modelo, incluindo piadas sobre Gary Marcus. As brincadeiras[35] encantaram os membros independentes do conselho; a apresentação descontraída também sinalizou, para eles, um avanço significativo nas capacidades da tecnologia e transmitiu a sensação de que os riscos envolvidos em suas decisões futuras estavam

aumentando.

Até então, Altman convocava o conselho cerca de uma vez por trimestre e preferia apresentar as atualizações verbalmente. Ele abordava tópicos de pesquisa complexos, às vezes trazendo pesquisadores da empresa para apresentar os avanços, e fazia resumos rápidos sobre os acordos que estava negociando com a Microsoft ou outros parceiros. Alguns dos membros independentes do conselho passaram a pressionar por reuniões mais frequentes e informações mais estruturadas, insistindo para que o executivo fornecesse relatórios escritos e lhes garantisse acesso a mais documentos.

Altman se irritou com o aumento da supervisão. Ele expressava repetidamente que queria que os integrantes do conselho atuassem mais como consultores, oferecendo sugestões para sua consideração. "O CEO deve tomar as decisões, e o conselho deve, você sabe, ser um fórum de discussão: conselhos e consentimento",[36] disse, certa vez, a estudantes universitários em 2017, e pareceu fazer referência a um conceito da Constituição dos Estados Unidos, essencial para o papel do Poder Legislativo de controlar o Executivo, mas usando a ideia para significar o oposto. "Eu acho que um conselho deve demitir um CEO ruim? Acho — e sei que isso é meio que uma heresia no Vale do Silício. Além disso, o conselho precisa dar ao CEO ampla liberdade para administrar a empresa? Sim."

Vários membros do conselho acreditavam que o conselho da OpenAI deveria ser diferente. "O conselho é um conselho sem fins lucrativos que foi criado com o objetivo de garantir que a missão de bem público da empresa fosse primordial, acima dos lucros, dos interesses dos investidores e de outras coisas", disse Helen Toner ao podcast *The TED AI Show*. "Não apenas, você sabe, ajudar o CEO a arrecadar mais dinheiro."[37]

Entre muitos funcionários, o GPT-4 fortaleceu a crença de que a IAG era possível. Pesquisadores antes céticos passaram a se sentir cada vez mais otimistas em relação à possibilidade de alcançar esse ápice técnico, mesmo enquanto a OpenAI continuava sem uma definição clara do que exatamente seria a IAG. Engenheiros e gerentes de produto que ingressaram na divisão de Produtos e tiveram a primeira interação próxima com IA por meio do GPT-4 adotaram uma linguagem ainda mais determinística. Para muitos colaboradores, a questão não era mais se a IAG aconteceria, mas quando.

Por outro lado, alguns funcionários sentiram exatamente o oposto. Embora houvesse uma mudança qualitativa evidente entre o que o GPT-2 e o GPT-3 conseguiam fazer, o GPT-4 era apenas maior, afirma um dos pesquisadores que trabalhou no modelo. "O grande resultado foi uma série de exames na qual o modelo se saía bem. Porém, mesmo isso é altamente questionável." A OpenAI nunca fez uma revisão abrangente dos dados de treinamento do GPT-4 para verificar se esses exames — e suas respostas — simplesmente presentes nos dados e sendo regurgitados, ou se o GPT-4 havia, de fato, desenvolvido uma nova capacidade que lhe permitia ser aprovado neles. Era o tipo de ciência frágil que se tornara comum com a mudança

generalizada do setor, da pesquisa revisada por pares para a pesquisa revisada pelo departamento de relações públicas.

Contudo, a crença de que a IA havia alcançado novos patamares estava no ar. Naquele primeiro semestre, no Google, Blake Lemoine, engenheiro da recém-reformada equipe de IA da gigante de tecnologia, estava convencido de que o grande modelo de linguagem da empresa, o LaMDA, não era apenas altamente inteligente, mas poderia ser considerado consciente. Ele afirmou que essa convicção não se baseava em uma avaliação científica, mas sim em sua crença, como padre cristão místico, de que Deus poderia escolher conceder consciência à tecnologia. "Quem sou eu para dizer a Deus onde ele pode ou não colocar almas?", escreveu Lemoine.[38] Quando os executivos do Google o demitiram, Lemoine tornou o caso público no *Washington Post*.[39] A repórter que divulgou a história, Nitasha Tiku, também conversou com Emily Bender e Meg Mitchell, que haviam alertado, no artigo "Papagaios Aleatórios", para o problema que era os grandes modelos de linguagem enganarem as pessoas, levando-as a enxergar significado e intenção reais onde havia apenas geração de texto.

"Agora temos máquinas que podem gerar palavras sem pensar, mas não aprendemos a parar de imaginar uma mente por trás delas", disse Bender.

"Estou realmente preocupada com o que significa as pessoas serem cada vez mais afetadas pela ilusão, especialmente quando essa ilusão está se tornando tão boa", afirmou Mitchell.

Nikhil Mishra, pesquisador de IA que estagiou na OpenAI em seus primeiros dias, traça paralelos com um experimento da década de 1970 que tentava ensinar a uma gorila uma forma modificada da língua de sinais dos Estados Unidos. Ao longo da vida, o animal, chamado Koko, pareceu aprender mais de mil sinais, e até desenvolveu a capacidade de construir frases. Contudo, apesar do enorme alarde público, outros especialistas argumentaram que Koko nunca adquiriu linguagem.[40] Embora fosse habilidosa em formar gestos, havia poucas evidências de que estivesse fazendo mais do que outros primatas haviam feito em experimentos semelhantes: simplesmente imitar os gestos dos cuidadores. Nenhum dado foi publicado sobre os sinais de Koko que permitisse verificar suas capacidades de forma independente, apenas vídeos selecionados dela demonstrando esses sinais. Às vezes, as apresentações ao vivo exigiam um olhar mais cuidadoso. Certa vez, quando o treinador perguntou a Koko se ela gostava de pessoas (*people*, em inglês), ela fez um gesto que significava "mamilo legal" (*fine nipple*). O treinador imediatamente explicou: *nipple* rimava com *people*; portanto, Koko achava que pessoas eram legais.

Para muitos observadores, esses episódios revelavam mais sobre a psicologia humana, e nossa tendência a projetar crenças e intenções onde não há nenhuma, do que sobre a capacidade de Koko. O treinador, diz Mishra, estava atribuindo significado onde não havia.

Para o próprio místico residente da OpenAI, Ilya Sutskever, que sempre acreditou mais do que a maioria dos pesquisadores na possibilidade de alcançar a IAG em curto prazo, o salto nas capacidades dos modelos da empresa serviu apenas como mais uma confirmação. Ele se convenceu de que estava testemunhando o surgimento de uma forma de raciocínio. Em conversas com Geoffrey Hinton,[41] Sutskever disse ao mentor que a IAG era iminente.

Antes, Sutskever concentrava-se sobretudo em incentivar os pesquisadores da OpenAI a desenvolverem novos recursos. Naquele momento, porém, passou a dedicar a atenção à pesquisa em segurança da IA com uma urgência renovada. Começou a repetir seu mantra "sinta a IAG" e incentivava as pessoas a se prepararem para mudanças drásticas.

"De repente, você vai se tornar a pessoa mais popular da festa", disse aos funcionários. "Você precisa não deixar isso subir à cabeça. Mantenha o foco na IAG, mantenha o foco na missão."

Naquele mês de setembro, a liderança técnica realizou um retiro no Tenaya Lodge, um resort de luxo remoto situado entre as exuberantes dobras da Serra Nevada. A propriedade, com interiores imponentes, várias piscinas e diversos restaurantes, ficava a apenas três quilômetros dos milhões de hectares de natureza intocada do Parque Nacional de Yosemite. Na primeira noite, todos se reuniram em torno de uma fogueira no pátio dos fundos do hotel. Cientistas seniores, vestidos com roupões de banho, formaram um semicírculo ao redor das chamas.

Então Sutskever surgiu. No centro da fogueira, ele havia colocado uma efígie de madeira encomendada a um artista local e iniciou uma performance dramática. A efígie, explicou, representava uma IAG boa e alinhada que a OpenAI teria construído, apenas para descobrir, depois, que ela era, na verdade, mentirosa e enganadora. O dever da OpenAI, comentava, era destruí-la. A poucos metros dali, várias sequoias erguiam-se como testemunhas antigas na escuridão. Sutskever despejou fluido de isqueiro sobre a efígie e ateou fogo.

CAPÍTULO 11

# ÁPICE

Em outubro de 2022, a OpenAI realizou um evento externo para toda a empresa em Monterey, na Califórnia, uma bela cidade costeira a duas horas ao sul de São Francisco. Naquele momento, a empresa havia crescido para cerca de trezentos funcionários, contra menos de duzentos no ano anterior. Para uma foto[1] tirada naquele fim de semana, todos posaram do lado de fora do pequeno aeroporto de Monterey, sorrindo e vestindo roupas com a marca da companhia. Altman parecia relaxado, sentado no chão à frente do grupo, com os joelhos contra o peito, os braços cruzados e os pés apontados para cima.

Durante dois dias, a equipe executiva apresentou atualizações sobre a visão e a execução: Altman falou sobre os enormes data centers que a Microsoft estava ampliando para a OpenAI; Steve Dowling e sua vice, Hannah Wong, detalharam as matérias de grande destaque que veículos de primeira linha estavam preparando sobre a empresa; Anna Makanju apresentou os avanços da presença cada vez maior da OpenAI na capital dos Estados Unidos, Washington. Em seguida, vieram as demonstrações, uma após a outra, dos projetos mais recentes das equipes de pesquisa e de produto. A grandiosidade de tudo aquilo era palpável. "Tudo junto era inacreditável", lembra um ex-funcionário que estava presente.

Brockman subiu ao palco para falar sobre os planos mais recentes para o GPT-4 e começou contando uma história sobre a esposa, Anna, que certo dia passou a sentir dores abdominais que não desapareciam. Ela era presença constante no escritório e tinha uma mesa ao lado da dele, onde, muitas vezes, o acompanhava em reuniões. Isso tudo sem trabalhar oficialmente na empresa. Além de Altman e Sutskever, ela era, como o próprio me disse uma vez, a pessoa em quem mais confiava para apoiá-lo, sua melhor amiga e confidente. Os dois costumavam estar grudados, indo a todos os lugares juntos. Greg contou que havia levado Anna a vários médicos, mas nenhum conseguiu identificar o problema. Quando ele perguntou ao GPT-4, porém, o modelo sugeriu uma doença que nenhum especialista havia considerado. "E ela tinha!", exclamou. Era uma releitura da mesma história que ele havia compartilhado comigo em 2019, sobre a promessa de que a IAG resolveria problemas de saúde de pessoas que consultaram inúmeros especialistas sem

respostas, desta vez com um personagem diferente.

Greg repetiria a história novamente semanas após a tentativa do conselho de demitir Altman, mas com uma nova variação. Ele descreveu os novos desafios médicos enfrentados por Anna e a luta ao longo de cinco anos, "consultando mais médicos e especialistas do que em toda a sua vida", até finalmente obter um diagnóstico. Foi um alergista que conseguiu juntar todas as peças e percebeu que ela tinha a síndrome de Ehlers-Danlos hipermóvel, um distúrbio genético que afeta a mobilidade. Ainda havia um longo caminho a percorrer antes que a IAG funcionasse "em áreas de alto risco como a medicina", escreveu ele, mantendo o tom firme em defesa da ambição da OpenAI. "Mas a promessa está ficando cada vez mais clara."

Poucas semanas após o retorno do evento externo, começaram a circular rumores de que a Anthropic estava testando — e em breve lançaria — um novo chatbot. A equipe do Superassistant estava na metade do desenvolvimento da interface de chat. Se não lançasse primeiro, a OpenAI corria o risco de perder a liderança, o que causaria um grande impacto no moral dos funcionários que haviam trabalhado longas e árduas horas para manter essa posição. E, pior ainda para alguns líderes, a OpenAI perderia para a Anthropic.

A Anthropic não estava, na verdade, planejando um lançamento iminente. Ela também enfrentava problemas. Com o colapso repentino da FTX no início de novembro, a empresa foi arrastada pelas consequências. Apenas alguns meses antes, havia levantado 580 milhões de dólares, dos quais 500 milhões, segundo um comunicado à imprensa da FTX, vinham de SBF e de outros integrantes da alta liderança. Documentos financeiros divulgados durante o julgamento revelariam mais tarde que esse dinheiro provinha dos bilhões que SBF havia desviado dos depósitos dos clientes da FTX,[2] tornando o investimento na Anthropic uma peça central no processo, especialmente na discussão sobre se os retornos significativos gerados por esse aporte poderiam ser usados para reembolsar os clientes. (Um juiz decidiria que sim,[3] e a FTX venderia as ações da Anthropic em lotes até 2024, totalizando 1,3 bilhão de dólares.)

No entanto, para os executivos da OpenAI, os rumores foram suficientes para desencadear uma decisão: a empresa não esperaria para preparar o GPT-4 como um chatbot. Em vez disso, lançaria o modelo GPT-3.5 com chat habilitado de Schulman, usando a nova interface de chat desenvolvida pela equipe do Superassistant. A equipe do Superassistant mudou instantaneamente de rumo, trazendo vários outros membros enquanto corria para integrar tudo e construir os recursos restantes. Para o restante da empresa, a liderança apresentou o esforço com cuidado. O ChatGPT — o nome escolhido — não seria, de fato, um lançamento de produto, mas sim uma "versão de teste", assim como com o DALL-E 2. Da mesma forma, não seria monetizado. O propósito seria "colocar o círculo de dados (flywheel) em

ação" — ou seja, acumular mais dados das pessoas que o utilizassem —, o que ajudaria a melhorar o GPT-4 e o produto Superassistant.

Fora da equipe do Superassistant, todos levaram os executivos ao pé da letra. Uma "versão de teste" não exigia atenção; era preciso se concentrar no lançamento do GPT-4 no início de 2023, e a equipe de Trust and Safety sentia que mal tinha tempo suficiente para construir a infraestrutura de monitoramento dentro desse prazo. Em comparação, o ChatGPT parecia não ser um problema. O GPT-3.5 já havia sido refinado com RLHF; era inerentemente mais seguro do que a versão do GPT-3, que, embora não tivesse sido lançada, estava disponível na API e não havia passado pelo mesmo refinamento. A OpenAI também havia publicado uma versão 3.5 sem recursos de chat em sua plataforma de desenvolvedores para que pudessem testar as capacidades do modelo. A adição de uma interface de chat realmente fazia diferença? As pessoas do clã de Segurança, ocupadas com os testes e ajustes do GPT-4, concordaram que não. Pela primeira vez, o lançamento de um modelo passou pelas verificações com pouca resistência.

Mesmo na equipe do Superassistant, ninguém compreendia de fato a mudança social que estava prestes a desencadear. Esperava-se que o chatbot fosse apenas um sucesso passageiro. Assim como o DALL-E 2, geraria muito alvoroço nas redes sociais, mas isso se acalmaria após algumas semanas. Na noite anterior ao lançamento, tudo parecia surpreendentemente calmo após uma corrida tão intensa até a linha de chegada. Eles fizeram apostas sobre quantos usuários experimentariam a ferramenta até o fim de semana, alguns estimando em alguns milhares; outros, dezenas de milhares. Por precaução, a equipe de infraestrutura provisionou capacidade de servidor suficiente para 100 mil usuários.

No dia seguinte, quarta-feira, 30 de novembro, a maioria dos outros funcionários nem percebeu que o lançamento havia ocorrido. Assim como a própria estreia da OpenAI, o lançamento do ChatGPT coincidiu com os procedimentos anuais do NeurIPS, realizado naquele ano em Nova Orleans, Louisiana. A conferência havia anunciado seu Test of Time Award [Prêmio Teste do Tempo], uma honraria concedida todo ano a um artigo publicado dez anos antes que tivesse tido impacto importante na área. Dessa vez, o ganhador tinha sido o artigo "ImageNet", de 2012, assinado por Hinton, Sutskever e Krizhevsky, trabalho que apresentou ao mundo o poder da aprendizagem profunda.

Naquela noite, um pequeno grupo de funcionários organizou uma festa da OpenAI perto do centro de convenções em que ocorreu a conferência para representar a empresa e recrutar candidatos entre os quase 10 mil participantes. O DeepMind, a Meta e o Google também realizaram festas de recrutamento concorrentes, espalhadas pela cidade no mesmo horário. Enquanto o evento transcorria, um recrutador notou que um engenheiro da OpenAI trabalhava sem parar em um computador.

Então, se ele aproximou do funcionário: "Cara, pega uma bebida. Está

todo mundo aqui. Relaxa um pouco, socializa."

O engenheiro nem levantou o olhar. "Não dá. Todas as GPUs estão derretendo. Está tudo travando."

Naquela noite, o Japão foi o primeiro a acordar, e imediatamente desencadeou um aumento maciço e inesperado no tráfego. No dia seguinte, o número de usuários continuou a disparar, fuso horário em fuso horário. Elon Musk ajudou a amplificar ainda mais essa onda. "Muitas pessoas presas em um maldito loop do ChatGPT ⇆", tuitou ele, recebendo cerca de 75 mil curtidas.

O sucesso instantâneo do ChatGPT ultrapassou qualquer expectativa dentro da OpenAI.[4] Isso deixaria os engenheiros e pesquisadores da empresa profundamente irritados, mesmo anos depois. O GPT-3.5 não representava uma grande evolução em relação ao GPT-3, que já estava no mercado havia dois anos, e o próprio GPT-3.5 já estava disponível para desenvolvedores. A interface e o formato do ChatGPT tornaram o modelo mais acessível, sem dúvida, mas não houve ali a transformação fundamental que muitos funcionários sentiram com o GPT-4. Altman diria mais tarde que acreditava que o ChatGPT seria popular, mas algo como "uma ordem de magnitude a menos".[5] "O fato de as pessoas gostarem foi chocante", recorda um ex-funcionário. "Para todos nós, eles pegaram algo inferior àquilo que usávamos internamente e lançaram."

Em cinco dias, Brockman postou no X que o ChatGPT havia ultrapassado um milhão de usuários. Em dois meses, já tinha alcançado 100 milhões, tornando-se, à época, o aplicativo para o consumidor final de crescimento mais rápido da história. (O rival Threads, da Meta, mais tarde reivindicaria esse título ao atingir o mesmo número de usuários em menos de cinco dias; especialistas argumentaram, porém, que isso não deveria contar, pois a Meta estava essencialmente aproveitando uma base já existente de usuários.)

O ChatGPT catapultou a OpenAI do status de startup em ascensão reconhecida dentro da indústria tecnológica para um nome familiar a todos da noite para o dia. Meses depois, em uma conferência de pesquisa em IA realizada em Kigali, Ruanda, a mais de 14 mil quilômetros de São Francisco, um pesquisador local comentou comigo que, após o lançamento do ChatGPT, até seus pais finalmente entenderam o que ele fazia no trabalho. "Você sabe que uma tecnologia é acessível quando a sua mãe comenta sobre ela", disse ele.

Ao mesmo tempo, foi esse sucesso estrondoso que colocou uma pressão extraordinária sobre a empresa. Ao longo do ano seguinte, ele ampliaria ainda mais a polarização entre suas facções e elevaria o estresse e a tensão interna a um nível explosivo.

Logo depois, toda a empresa entrou em modo de emergência. Os servidores da OpenAI travaram repetidas vezes enquanto a equipe de

infraestrutura tentava ampliar a capacidade no prazo mais apertado da história do Vale do Silício. Parte da computação destinada à divisão de Pesquisa foi temporariamente desviada para sustentar o crescimento explosivo do ChatGPT, e, ainda assim, não era suficiente para manter o aplicativo estável. A equipe de Trust and Safety, com pouco mais de uma dúzia de pessoas,[6] lutava para identificar e conter comportamentos abusivos em meio à enxurrada de novos usuários, prejudicada por sistemas de monitoramento frágeis. Ela também enfrentava dificuldade para contratar os engenheiros necessários para implementar o plano de fiscalização reativa, ainda em desenvolvimento. O ChatGPT interrompeu tudo. Os esforços para finalizar as novas ferramentas foram suspensos, conforme os recursos de engenharia eram redirecionados para manter o serviço de pé. E, quando os servidores caíam, a própria plataforma de monitoramento de tráfego caía junto, tornando impossível aplicar qualquer medida de moderação em escala.

A grave escassez de GPUs também prejudicou outro esforço.[7] Na tentativa de aproveitar a tecnologia da própria empresa,[8] a equipe de Trust and Safety criou o protótipo de um plano interno chamado Fact Factory, que a OpenAI divulgou publicamente, para usar o GPT-4 na moderação do conteúdo gerado por seus próprios modelos e por outros modelos da OpenAI. A implementação, porém, não era exatamente escalonável: ela exigia fornecer ao GPT-4 prompts extremamente longos para capturar nuances suficientes. Mesmo quando os servidores estavam funcionando, isso consumia muitos recursos computacionais, e os servidores não estavam sempre disponíveis.

Para muitos no clã de Segurança, o ChatGPT tinha sido até então o exemplo mais alarmante das limitações da capacidade preditiva da OpenAI. Um integrante da equipe da área levantou a questão em uma reunião com todos os funcionários: como a empresa poderia ter falhado tão gravemente em prever o comportamento dos usuários e a popularidade do ChatGPT? O que isso dizia sobre a habilidade da própria OpenAI de calibrar e antecipar os impactos futuros de suas tecnologias?

Para grande parte do restante da empresa, os servidores em colapso, embora fossem uma fonte extraordinária de estresse, eram, também, uma marca extraordinária de triunfo. A OpenAI havia criado uma tecnologia tão profunda, e com uma demanda tão grande, que agitou o mundo e o transformou da noite para o dia. Eles tinham seus olhos voltados para toda a humanidade, e realmente haviam conseguido. Todos os 8 bilhões de pessoas passaram a morar no mundo da OpenAI.

Altman não se deixou levar pelo momento. Ele lembrou aos funcionários que a empresa tinha, em última análise, a missão de alcançar algo muito maior do que construir "o maior produto da história do Vale do Silício". Ele incentivou todas as equipes a manterem o rumo e seguirem em frente. Como esperava,[9] a OpenAI havia despertado todos os concorrentes: a Anthropic estava prestes a lançar seu chatbot, Claude; o Google havia soado um alarme interno de "código vermelho" e logo consolidaria suas divisões de IA no

Google DeepMind para concentrar todo o peso no lançamento de um produto semelhante. Embora a OpenAI tivesse chegado primeiro ao mercado com uma oferta dez vezes melhor, ela precisava continuar correndo para permanecer na liderança.

Com todas as equipes perigosamente sobrecarregadas, os gerentes imploraram a Altman por mais pessoal, e candidatos não faltavam. Depois do ChatGPT, o número de interessados em trabalhar na empresa multiplicou-se rapidamente. No entanto, Altman se preocupava com o impacto que uma expansão acelerada poderia ter sobre a cultura organizacional e o alinhamento com a missão. Ele acreditava em manter uma equipe pequena e de alta densidade de talentos. "Estamos agora em uma posição na qual é tentador deixar a organização crescer com força",[10] escreveu em seu memorando de visão de 2020, referindo-se ao investimento da Microsoft. "Devemos nos esforçar muito para resistir a isso — o que tem funcionado para nós até o momento é sermos pequenos, focados, altamente confiáveis, intensos e sem besteiras."

"A sobrecarga de pessoal e a burocracia excessiva podem facilmente matar grandes ideias ou resultar em esclerose. Ao contrário dos grandes projetos de engenharia do passado, conseguimos cumprir nossa missão com um número surpreendentemente pequeno de pessoas excelentes."

Ele repetia isso aos executivos no fim de 2022, enfatizando de forma contínua, durante as discussões sobre o número de funcionários, a necessidade de manter a empresa enxuta e o nível de talento alto, e de não contratar mais do que cem pessoas. Outros executivos resistiram. Com o ritmo em que suas equipes estavam ficando esgotadas, muitos acreditavam ser necessário algo mais próximo de quinhentas, ou até mais contratações.

Ao longo de várias semanas, conforme as discussões avançavam, a equipe executiva finalmente chegou a um acordo sobre um número intermediário, algo entre duzentos e cinquenta e trezentos funcionários. Contudo, esse limite não se sustentou. No meio do ano, havia cerca de trinta, ou cinquenta, novas contratações por semana, incluindo mais recrutadores para acelerar ainda mais o processo. No meio do segundo semestre, a empresa já havia ultrapassado em muito a cota que ela própria havia estabelecido.

O repentino surto de crescimento transformou a cultura da empresa. Um recrutador escreveu um manifesto afirmando que a pressão para contratar tão rapidamente estava obrigando sua equipe a reduzir o padrão de qualidade dos candidatos. "Se você quer construir a Meta, está fazendo um ótimo trabalho", disse ele, em uma crítica direta a Altman, e aludindo aos temores que o CEO havia expressado sobre a empresa diluir a densidade de talentos e a orientação para a missão enquanto ampliava a burocracia. A expansão acelerada também provocava um aumento nas demissões. Durante a integração, um gerente foi instruído a documentar e relatar qualquer membro de sua equipe com baixo desempenho, apenas para ele mesmo ser demitido algum tempo depois. As saídas raramente eram comunicadas ao restante da empresa. As pessoas

descobriam que colegas haviam sido dispensados apenas ao notar que uma conta no Slack havia ficado cinza por ter sido desativada. Passaram a chamar isso de "desaparecer".

Para os novos contratados, totalmente convencidos de que estavam entrando em uma startup dinâmica e lucrativa, a turbulência parecia uma manifestação particularmente caótica, às vezes brutal, de problemas corporativos comuns: gestão deficiente, prioridades confusas e a crueldade impiedosa de uma organização capitalista disposta a tratar os funcionários como descartáveis. "Havia uma enorme falta de segurança psicológica", diz um ex-funcionário que ingressou na época. "É como o oposto de 'uma empresa como uma família', o que é justo, você sabe, é uma empresa." Muitas pessoas que entraram nesse período relatam essa mesma falta de segurança psicológica. Alguns se sentiam como se estivessem em uma prisão, agarrando-se com unhas e dentes apenas para completar um ano e ter acesso à primeira parcela das ações. Ainda assim, havia uma vantagem significativa: muitos sentiam que os colegas estavam entre os melhores do setor de tecnologia, o que, combinado aos recursos aparentemente ilimitados e ao impacto global incomparável, podia criar uma sensação quase mágica quando tudo realmente se alinhava. "Eu diria que a OpenAI é um dos melhores lugares onde já trabalhei, mas também provavelmente um dos piores", resume o ex-funcionário.

Para alguns trabalhadores que se lembravam dos primeiros dias difíceis da OpenAI, quando era uma entidade sem fins lucrativos, coesa e orientada para a missão, a transformação drástica em uma grande corporação sem rosto foi chocante e emocionante. A organização que conheciam havia desaparecido; no lugar surgiu algo irreconhecível. "A OpenAI é o Burning Man", diz Rob Mallery, ex-recrutador, referindo-se à forma como o festival de arte no deserto cresceu a tal ponto que perdeu contato com sua essência. "Sei que significava muito mais para as pessoas que estavam lá no início do que significa para todos agora."

Naqueles primeiros anos, a equipe criou um canal no Slack chamado #explainlikeimfive [#meexpliquecomoseeutivessecincoanos], que permitia aos funcionários enviar perguntas anônimas sobre temas técnicos. Com o número de contratados da empresa chegando a seiscentos, o canal também passou a servir como espaço para reclamações anônimas. Em meados de 2023, alguém postou que a corporação estava contratando muitas pessoas que não estavam alinhadas com a missão ou apaixonadas pela construção da IAG.

Outra pessoa respondeu: sabiam que a OpenAI entraria em declínio assim que começasse a contratar pessoas que conseguiam olhar você nos olhos.

O ChatGPT também surpreendeu a Microsoft. Os líderes da OpenAI haviam dito à parceira, assim como aos próprios funcionários, que o chatbot seria um "versão de teste". Claramente, foi tudo menos isso.

A incompatibilidade inicialmente irritou os executivos da gigante da tecnologia. O ChatGPT roubou completamente a cena do chatbot da

Microsoft para o Bing. Quando ela lançou o Bing AI em fevereiro seguinte, o produto sofreu um golpe de relações públicas com um artigo do colunista do *New York Times* Kevin Roose, que relatava que o chatbot o havia levado a considerar divorciar-se da esposa. Isso estava longe da recepção que a Microsoft esperava e, em comparação, fez a OpenAI parecer ainda melhor.

No entanto, os mal-entendidos não foram suficientes para diminuir o entusiasmo da Microsoft pela OpenAI. A recepção extremamente positiva do ChatGPT foi contagiante, e as capacidades cada vez maiores dos modelos da OpenAI deixaram os executivos da gigante ainda mais entusiasmados. A Microsoft estava preparando uma linha totalmente nova de Copilots para seus produtos, baseada no GPT-3.5 e no GPT-4, e planejava lançar um após o outro em uma série constante de anúncios. Depois do Bing em fevereiro, março foi dedicado ao Microsoft 365 Copilot, que levaria uma interface baseada em chat e alimentada por IA a todos os produtos do Office, do Word ao Outlook e ao Teams.

A maneira como os executivos da Microsoft falavam internamente sobre a parceria com a OpenAI estava mudando depressa.[11] Antes, a gigante sentia que tinha poder sobre a OpenAI; naquele momento, alguns executivos percebiam que era a OpenAI que tinha poder sobre a Microsoft. Havia, em setores da empresa, uma sensação crescente de inadequação diante do fato de que os próprios esforços de pesquisa em IA não haviam alcançado o que a OpenAI conseguira. Se a Microsoft se afastasse como principal investidora da empresa menor, a startup poderia encontrar outros financiadores, lembra um ex-funcionário da Microsoft sobre o pensamento de alguns executivos. Contudo, se a OpenAI se afastasse da Microsoft, a gigante encontraria outra OpenAI?

Ao mesmo tempo, muitos executivos já não falavam apenas em derrotar o Google. Antes, eles reagiam às declarações estranhas da OpenAI sobre a IAG e à linguagem pomposa sobre seu poder de inventar o futuro com um ceticismo educado, mas depois passaram a acreditar. IA, IAG, IA generativa — como quer que se chamasse — era o futuro, e a Microsoft estava conduzindo esse futuro de mãos dadas com a OpenAI. Quanto mais ela acreditava, mais reorientava sua retórica e estratégia. "Vi os incentivos na Microsoft levarem cada vez mais a uma concepção estreita do futuro", diz o ex-funcionário. "Vi a tecnologia tornar-se algo restrito, sustentado pela narrativa, em vez da realidade."

Satya Nadella implementou uma nova estratégia para distribuir os recursos de computação da Microsoft.[12] Ele retirou as GPUs das equipes de pesquisa da própria empresa para apoiar a OpenAI, consolidando, também, todas as GPUs em um único pool para sustentar melhor as cargas de trabalho de IA generativa. "O típico funcionário da Microsoft não tinha a menor ideia do que era a OpenAI antes de janeiro do ano passado", recorda um empregado da Microsoft. Naquele momento, esses funcionários estavam recebendo diretrizes urgentes dos superiores sobre como encontrar maneiras de integrar o trabalho

às tecnologias da OpenAI.

A gigante da tecnologia experimentaria uma rápida proliferação de mais de cem novos projetos de IAG em apenas alguns meses, à medida que os funcionários testavam diversas maneiras de usar o GPT-4 e o ChatGPT. Em uma reviravolta irônica, essa adoção agressiva forçaria a Microsoft a lidar com muitos dos mesmos desafios que outras empresas enfrentariam ao correr para adotar a IA generativa sem compreendê-la totalmente. Isso incluía causar dores de cabeça para as equipes de Risco e Conformidade. Nem todos estavam usando as versões internas das tecnologias da Microsoft; alguns optavam por usar a versão gratuita do ChatGPT diretamente da OpenAI, que treinava com dados do usuário, levantando preocupações sobre um possível vazamento de informações de clientes da Microsoft ou uma interferência na conformidade regulatória. Embora alguns funcionários considerassem as ferramentas um grande impulso à produtividade, muitos as achavam exaustivas. "Existe essa cultura de 'use IA, use IA, use IA'", diz um deles. "Mas é como se isso não nos ajudasse. Não queremos usá-la, mas parece que ela está em toda parte e não podemos escapar dela."

Ao mudar dos modelos próprios da Microsoft para os da OpenAI, muitos funcionários também perderam o controle e a visibilidade de uma camada essencial da infraestrutura de seu trabalho. Dentro da gigante da tecnologia, o acesso aos modelos subjacentes da startup era rigidamente protegido, mesmo que fossem treinados e armazenados nos servidores da Microsoft. A maioria dos funcionários da empresa já não podia examinar os dados de treinamento nem ajustar os pesos dos modelos que utilizavam. Os modelos da OpenAI eram fornecidos por meio de uma API, assim como para outros clientes da OpenAI.

Contudo, em troca dessas concessões, a Microsoft estava sendo amplamente recompensada. A empresa observava um aumento maciço no número de clientes para a plataforma Azure AI, por ser o único provedor de nuvem capaz de oferecer os benefícios típicos da nuvem, como armazenamento e gerenciamento de dados mais simples, além da capacidade de processar esses dados com os recursos da OpenAI. "O Azure OpenAI Service está nos abrindo as portas para muitos clientes atualmente",[13] escreveu Eric Boyd, vice-presidente corporativo da plataforma de IA, em um e-mail para sua divisão em maio de 2023. Em agosto, ele voltou a se entusiasmar. "De vez em quando, é ótimo olhar para trás e se maravilhar com o quanto avançamos em apenas um ano", escreveu, acrescentando que a plataforma havia registrado naquele ano um "aumento de 21 vezes no número de clientes". No mês seguinte, Boyd comemorou um novo marco. Após centralizar os esforços fragmentados de IA da Microsoft na plataforma e com os milhares de novos clientes que aderiram ao Azure OpenAI Service, o tráfego na plataforma havia crescido dez vezes em apenas nove meses. Em janeiro de 2023,[14] ela recebia 100 bilhões de pedidos de inferência mensais; já em setembro, recebia um trilhão.

Naquele verão, o diretor de tecnologia da Microsoft, Kevin Scott, foi efusivo nos elogios durante uma reunião geral da OpenAI. "Paramos nossos investimentos em aprendizagem de máquina de IA em vários lugares, a ponto de as pessoas ficarem tipo: 'Ei, sabe de uma coisa, Microsoft, vá se danar'",[15] disse Scott, referindo-se à retirada de GPUs de algumas pesquisas internas da própria Microsoft. "E nós aceitamos o desafio porque acreditamos que vocês estão fazendo o melhor trabalho do setor."

O ChatGPT consolidou a transição definitiva da OpenAI de uma organização sem fins lucrativos para uma voltada para a comercialização. Altman e outros executivos pressionaram para aproveitar o impulso do sucesso do chatbot, lançando uma série de produtos pagos. Em fevereiro de 2023, foi lançada uma versão paga do ChatGPT; em março, vieram, uma após a outra, a versão em API, a Whisper API e, finalmente, o GPT-4. "Depois do ChatGPT, havia um caminho claro para receita e lucro", diz um ex-funcionário. "Não era mais possível defender a ideia de ser um laboratório de pesquisa idealista. Havia clientes que queriam ser atendidos na hora."

A explosão de novos produtos sobrecarregou de novo a equipe de Trust and Safety. Por um tempo, a OpenAI atraiu usuários para aderir à sua API oferecendo um crédito inicial de vinte dólares em uso. Com um aumento drástico no uso da API provocado pela fama do ChatGPT, esse incentivo de inscrição passou a representar um problema: muitos usuários estavam criando contas em grande escala para lucrar repetidamente com o bônus. Em alguns casos, também criavam contas adicionais para contornar banimentos e suspensões nas anteriores. A fraude em massa estava fazendo a OpenAI perder enormes quantias de receita, à medida que os custos aumentavam com a necessidade de mais servidores. Ainda com menos de vinte pessoas e com os esforços de fiscalização reativa severamente prejudicados,[16] a equipe de Trust and Safety redirecionou seu pessoal para, mais uma vez, combater o novo vetor de abuso.

Em pouco tempo, as constantes mudanças levariam Willner a um grave quadro de burnout. Em poucos meses, ele e vários de seus funcionários deixariam a empresa. No fim daquele ano,[17] a equipe seria dissolvida, e alguns dos integrantes remanescentes seriam incorporados a uma operação mais ampla de sistemas de segurança, liderada por Lilian Weng, pesquisadora de longa data da OpenAI. Alguns da equipe de Trust and Safety passaram a achar que o conflito persistente entre as divisões de Produtos e Segurança, com sua ênfase exagerada no pessimismo, havia cultivado, entre muitos líderes da empresa, uma cultura de forte desconsideração por qualquer tipo de "segurança". Para esses funcionários, o ambiente criado pelo embate entre as duas áreas acabou tornando sua função, já fragilizada na maioria das empresas de tecnologia, ainda mais enfraquecida dentro da OpenAI.

De fato, a cada novo lançamento, o conflito continuava — inclusive em relação ao lançamento do GPT-4. Embora muitos na equipe de Produtos

considerassem o atraso de seis meses no lançamento do modelo uma medida excessivamente cautelosa, alguns na equipe de Segurança achavam que ainda não havia tempo suficiente para concluir os testes abrangentes e o alinhamento.

A alta taxa de alucinações do modelo, por exemplo, permanecia particularmente difícil de controlar, mesmo com um esforço conjunto de RLHF para mitigar o problema. Em novembro de 2022, enquanto os usuários recorriam ao ChatGPT como se fosse uma ferramenta de pesquisa[18] e geravam especulações generalizadas de que ele poderia destronar o Google, um documento interno observou que o modelo da OpenAI havia alucinado, durante um teste, em cerca de 30% das chamadas perguntas de domínio fechado.

As perguntas de domínio fechado são consideradas a categoria mais fácil de pergunta: são aquelas em que os usuários pedem ao modelo que trabalhe apenas com as informações fornecidas — por exemplo, ao carregar um PDF e solicitar um resumo, ou ao fornecer pontos-chave e pedir que o modelo complete ou reescreva frases. Isso contrasta com as perguntas de domínio aberto, quando o usuário faz uma pergunta sem fornecer material de referência — sobre cultura pop, história antiga, biologia do ensino médio —, da mesma forma que faria em um mecanismo de pesquisa típico.

Enquanto isso, as GPUs se tornaram uma restrição constante para a pesquisa e a expansão da OpenAI. Novos estudos, bem como lançamentos de produtos ou recursos, precisaram ser repetidamente adiados ou arquivados devido à falta de capacidade dos chips. Depois que o ChatGPT viralizou, a *SemiAnalysis*, um boletim informativo comercial focado na indústria de semicondutores, estimou que a empresa estava gastando cerca de 700 mil dólares por dia apenas em custos de computação.[19] Após os executivos realocarem chips da divisão de Pesquisa, a divisão de Produtos se comprometeu a devolvê-los até determinada data. Essa data chegou e passou, mas a divisão de Produtos não conseguiu devolver os chips. Com crescimento rápido do número de usuários, a divisão precisava de mais chips, não de menos.

A pressão acelerou a pesquisa da OpenAI a fim de desenvolver modelos mais eficientes. Durante o trabalho para melhorar a eficiência computacional da empresa, a divisão de Pesquisa descobriu um novo método para desenvolver modelos baseados em Transformer mais baratos para servir os usuários. Ao usar esse método, que chamaram de DUST, eles atribuíram nomes de código aos modelos resultantes seguindo um tema inspirado no deserto. O primeiro deles, uma versão otimizada do GPT-3.5, foi chamado de Saara, lançado em fevereiro de 2023 com o nome público GPT-3.5 Turbo. Outro recebeu o nome de Gobi, e seria uma versão otimizada de um de seus modelos de texto e imagem.

Um terceiro modelo, que deveria ser uma versão otimizada do GPT-4, recebeu o nome de Arrakis, o planeta deserto da épica ficção científica *Duna*.

Contudo, após meses de trabalho, a equipe ainda enfrentava dificuldades para tornar o Arrakis mais eficiente sem perder desempenho. O projeto consumia recursos computacionais significativos. Pouco tempo depois, a liderança decidiu descartá-lo para liberar GPUs para outros projetos.

Enquanto a Microsoft se preocupava com a possibilidade de a OpenAI abandonar a parceria, a OpenAI sentia a própria vulnerabilidade diante da chance de a Microsoft deixar de cooperar caso a startup não se esforçasse para agradar a parceira. O Arrakis foi visto como um revés particular nesse sentido. [20] Na esperança de impressionar a gigante, a OpenAI reformulou o plano de ação para priorizar a entrega do modelo, em detrimento dos próprios projetos mais alinhados estrategicamente, incluindo uma iniciativa para aplicar o GPT-4 a um produto de mecanismo de busca. Entretanto, em vez disso, a iniciativa fracassada deixou alguns executivos seniores da Microsoft decepcionados.

Havia também uma nova e incômoda realidade: a OpenAI e a Microsoft estavam começando a competir por contratos.[21] O Codex e o DALL-E 2 haviam convencido a OpenAI a manter o controle da entrega de suas tecnologias diretamente aos usuários. O ChatGPT e o GPT-4 mostravam que a OpenAI também podia gerar a própria receita. Isso significava oferecer diretamente aos clientes a mesma tecnologia que fornecia à Microsoft, tecnologia que a Microsoft, por sua vez, oferecia aos mesmos clientes.

A transferência da tecnologia tinha os próprios desafios. À medida que o cronograma de lançamentos da OpenAI se acelerava, o mesmo acontecia com o da Microsoft. Contudo, a Microsoft acabava ofuscando a OpenAI, criando uma dinâmica em que um único funcionário da OpenAI podia ser contatado por dezenas de colegas da Microsoft, em vários departamentos, com todos os tipos de perguntas sobre detalhes técnicos ou logísticos de cada novo produto. Isso estava se tornando cada vez mais frustrante e opressivo para a equipe da OpenAI, que precisava dar suporte aos lançamentos da Microsoft enquanto tentava se concentrar no próprio plano de trabalho.

Grande parte do trabalho de amenizar o relacionamento ficou a cargo de Murati. Ela procurou trabalhar em estreita colaboração com Scott e outros executivos da Microsoft para coordenar o calendário de lançamentos de produtos, traçar estratégias para diferenciar as ofertas da OpenAI e da Microsoft e encontrar maneiras mais produtivas para as duas organizações colaborarem. No verão de 2023, numa tentativa de reduzir a carga de comunicação, uma equipe de engenheiros da Microsoft passou a se integrar à OpenAI, com acesso total a tudo, para otimizar as transferências de tecnologia.

Com a integração mais profunda, Altman e Nadella estavam se envolvendo mais do que nunca, especialmente na gestão de recursos computacionais, tomando decisões difíceis sobre como redistribuir chips e dinheiro para ainda mais chips para as operações da OpenAI, sempre famintas

por computação. Nadella disse ao *New York Times* que as demandas da OpenAI cresciam tão rápido e eram tão altas que Altman começou a ligar para ele todos os dias dizendo: "Preciso de mais, preciso de mais, preciso de mais."[22]

A OpenAI não precisava apenas de mais data centers para manter o ChatGPT. Ela precisava de supercomputadores muito mais potentes para treinar as futuras gerações de modelos. Para atender a essa demanda agressiva e crescente,[23] as duas organizações estavam esboçando um novo projeto sem precedentes chamado Stargate para a OpenAI, e Mercury para a Microsoft: um único supercomputador que, apenas para ser construído, custaria cerca de 100 bilhões de dólares.

O império da IA estava voltando à mesma forma de expansão dos impérios antigos: para alimentar o crescimento, precisava de mais recursos materiais e, fundamentalmente, de mais terras.

CAPÍTULO 12

# TERRA SAQUEADA

Em Santiago, no Chile, nos dias que se seguem a uma boa chuva, a poluição se dissipa e revela uma vista deslumbrante da cordilheira dos Andes. A cadeia montanhosa percorre toda a extensão desse país estreito, da Patagônia, no extremo sul, até a fronteira norte, por mais de 4.200 quilômetros o dobro da distância entre Miami e Boston. Nessa região, as paisagens urbanas da capital dão lugar a uma vastidão de deserto que, com exceção de partes da Antártida, é o lugar mais seco da Terra. As montanhas tornam-se mais nítidas, as formas mais definidas, e as cores ganham intensidade sob o sol forte.[1]

Antes de o Chile ser o Chile, o deserto do Atacama, como passou a ser conhecido, era o lar de diversos grupos indígenas. Onde muitos enxergavam uma paisagem árida e inóspita, eles fizeram do deserto um lugar habitável, extraindo água e minerais das profundezas da terra para cultivar, criar gado e realizar rituais ancestrais. Então chegaram os espanhóis. Eles dividiram a região em unidades administrativas com fronteiras definidas. Até hoje, anciãos indígenas ainda contam histórias sobre uma repressão tão violenta que os espanhóis cortavam a língua de quem ousasse continuar falando o idioma nativo. A partir desse momento, o império moldou a relação do país com o restante do mundo: o Chile forneceria recursos brutos — terra, água, energia, minerais — para sustentar os planos políticos e econômicos de outras nações.[2]

Hoje, quase 60% das exportações do Chile são minerais,[3] extraídos principalmente no deserto do Atacama, sobretudo cobre, um metal altamente condutor usado em todos os tipos de eletrônicos, e, mais recentemente, lítio, o ingrediente essencial das baterias de íon-lítio. Essas e outras exportações de recursos impulsionam a economia do país. Em Santiago, todos conhecem alguém que vive no ritmo da indústria de mineração: durante seus "turnos" de trabalho, mora no norte; nos dias de folga, retorna à capital.

O país tem enfrentado dificuldades para desenvolver qualquer outra indústria que funcione como motor econômico.[4] Muito tempo após o fim do império espanhol, os Estados Unidos desempenharam um papel fundamental para manter essa trajetória.[5] Nos anos 1950 e 1960,[6] à medida que a

economia desenvolvimentista se enraizava no Chile e no Uruguai, favorecendo uma forte regulamentação governamental e um foco interno na industrialização como caminho para o amadurecimento de economias em desenvolvimento, as multinacionais dos Estados Unidos que haviam ganhado bilhões com suas participações nas minas chilenas começaram a se irritar com os crescentes impostos e as restrições estatais. O governo norte-americano então embarcou em uma iniciativa para reformular as políticas econômicas do Chile de modo a torná-las mais favoráveis a seus interesses comerciais, lançando em 1956 o "Projeto Chile", um programa para educar cem estudantes chilenos na Universidade de Chicago sob a tutela intelectual do economista Milton Friedman.

Friedman foi uma figura proeminente na economia, que viria a receber o Prêmio Nobel em 1976,[7] e cujas ideias podem ser resumidas pelo título de seu influente artigo publicado em 1970 no *New York Times*: "A Friedman Doctrine — The Social Responsibility of Business Is to Increase Its Profits" [A responsabilidade social das empresas é aumentar seus lucros, em tradução livre]. Ele defendia tudo o que a economia desenvolvimentista não defendia: nenhuma regulamentação governamental, liberdade irrestrita para empresas com fins lucrativos e um caminho para a maturação econômica dos países em desenvolvimento baseado no enfrentamento externo, como por meio de exportações liberais. Como Naomi Klein detalha em seu best-seller internacional de 2007, *A doutrina do choque*,[8] o Projeto Chile não era educação, mas doutrinação. Na Universidade de Chicago, estudantes chilenos — e, mais tarde, estudantes de outros países da América Latina — eram explicitamente ensinados a criticar as políticas econômicas de seus países e as falhas fatais do desenvolvimentismo latino-americano.

À medida que cada turma de formandos, conhecida como "Chicago Boys", retornava a Santiago, as ideias neoliberais de Friedman foram se espalhando pela elite intelectual do Chile, até se tornarem parte da ideologia dominante. Em 1973, o presidente de esquerda democraticamente eleito do Chile foi derrubado pelo general Augusto Pinochet em um golpe de Estado fomentado, em parte, pela CIA.[9] O golpe marcou o início de quase duas décadas da ditadura brutal de Pinochet e de uma nova agenda econômica neoliberal: a ditadura nomeou os Chicago Boys para elaborarem suas políticas econômicas.

Sob o regime de Pinochet,[10] o Chile privatizou praticamente tudo — educação, saúde, sistema previdenciário e até o abastecimento de água. A estratégia gerou crescimento econômico, mas também alimentou uma desigualdade impressionante. Hoje, o Chile está entre os países mais desiguais do mundo, com quase um quarto da renda nacional concentrada em algumas poucas famílias que representam 1% da população.[11] Como nunca passou por uma industrialização significativa, o país também permanece preso a uma economia extrativista, o que o mantém relevante para potências geopolíticas superiores.

Assim, com o boom da IA, o Chile se tornaria o ponto zero de uma nova etapa de extrativismo, como fornecedor para o apetite insaciável da indústria por recursos brutos, não apenas o cobre e lítio no norte, mas também seus recursos terrestres, hídricos e energéticos para um número crescente de data centers na região metropolitana de Santiago. Em maio de 2024, o governo anunciou com orgulho que o país receberia 28 novos data centers, além dos 22 já existentes, nos próximos anos, atraindo 2,6 bilhões de dólares em investimentos estrangeiros.[12]

Embora o governo sustente que seu papel como fornecedor de recursos para o desenvolvimento tecnológico representa progresso para o país, o Chile também se tornou palco de algumas das resistências mais intensas do mundo contra essa narrativa. Comunidades em todo o país estão lutando contra a desapropriação de terras, de água e de outros recursos a serviço de visões do Norte Global que não as incluem nem as beneficiam. Por meio de protestos de rua e batalhas judiciais, os esforços atrasaram projetos de empresas e chamaram a atenção do governo, além de inspirarem pessoas em outros países a se mobilizarem em solidariedade.

Martín Tironi Rodó, professor da Universidade Católica de Santiago e diretor do centro de pesquisa chileno Futures of Artificial Intelligence Research [Futuros da Pesquisa em Inteligência Artificial], que estuda IA a partir de uma perspectiva interdisciplinar e latino-americana, resume o sentimento que ouvi repetidamente dessas comunidades enquanto viajava pelo país para encontrá-las. A questão central que esses movimentos estão levantando é como imaginar um caminho diferente para o desenvolvimento da IA que não esteja enraizado na extração, diz ele. "Se formos desenvolver essa tecnologia da mesma forma que fazíamos, vamos devastar a Terra."[13]

As tecnologias "digitais" não existem apenas digitalmente. A "nuvem" não assume de fato a forma etérea que seu nome sugere. Para treinar e disponibilizar modelos de IA, são necessários data centers físicos e tangíveis. E, para treinar e executar os tipos de modelos de IA generativa que a pioneira OpenAI desenvolveu, são necessários data centers maiores e em maior número do que nunca.

Antes da IA, os data centers já estavam crescendo e se expandindo. Eles já foram pequenos e distribuídos de forma que ficavam ocultos em ambientes urbanos, algumas prateleiras de computadores escondidas em um armário de escritório ou algumas dezenas de racks em um prédio adaptado. Nos anos 2000, os gigantes da tecnologia começaram a seguir uma direção diferente, consolidando toda a infraestrutura de computação em enormes armazéns de servidores situados em comunidades rurais. O mundo dos data centers ficou dividido: havia os hiperescaladores e havia todos os outros. Os quatro maiores hiperescaladores — Google, Microsoft, Amazon e Meta — gastam mais dinheiro construindo data centers a cada ano do que praticamente todos os outros, incluindo desenvolvedores relativamente desconhecidos como Equinix

e Digital Realty, combinados.[14]

É difícil imaginar como é um data center de hiperescala se você nunca viu um.[15] Mél Hogan, professora associada da Queen's University, no Canadá, que estuda IA, infraestrutura e meio ambiente, costumava usar campos de futebol para descrevê-los quando começou a escrever sobre eles há cerca de uma década. "Agora, os campos de futebol nem chegam perto do tamanho necessário",[16] diz ela. Os hiperescaladores chamam seus data centers de "campi" — grandes extensões de terra que rivalizam com as maiores universidades da Ivy League, com vários edifícios enormes, densamente lotados de racks e mais racks de computadores. Essas máquinas emitem uma quantidade imensa de calor, como um laptop em pleno funcionamento multiplicado por um milhão. Para evitar o superaquecimento, os edifícios também contam com sistemas de refrigeração gigantescos — grandes ventiladores, aparelhos de ar-condicionado ou sistemas que evaporam água para resfriar os servidores. Todo esse equipamento cria uma cacofonia de zumbidos, ruídos e estalos que pode — especialmente em comunidades pouco desenvolvidas — ser ouvida a quilômetros de distância, 24 horas por dia, tornando-se uma fonte constante e perturbadora de poluição sonora.[17]

Hoje, os desenvolvedores usam uma nova palavra para distinguir a escala do que está por vir na era pós-ChatGPT: *megacampus*.[18] O termo se refere não apenas à área de terreno, mas também à enorme quantidade de energia necessária para operá-los. Um rack de GPUs consome três vezes mais energia do que um rack de outros tipos de chips.[19] E não é apenas o treinamento dos modelos de IA generativa que é caro, o próprio funcionamento desses modelos demanda muito. De acordo com a Agência Internacional de Energia dos Estados Unidos, cada consulta ao ChatGPT consome, em média, cerca de dez vezes mais eletricidade do que uma pesquisa típica no Google.[20]

Até recentemente, os maiores data centers eram projetados para operar com cerca de 150 megawatts, o que significa consumir anualmente tanta energia quanto aproximadamente 122 mil residências nos Estados Unidos.[21] Hoje, desenvolvedores e empresas de serviços públicos se preparam para megacampi de IA que poderão exigir de mil a 2 mil megawatts. Um único deles poderia usar, em um ano, de cerca de uma vez e meia a três vezes e meia a quantidade de energia consumida pela cidade de São Francisco.[22]

Existem poucos lugares no planeta capazes de produzir e fornecer tanta energia para um único local. Os desenvolvedores estão trabalhando com empresas de serviços públicos ao redor do mundo para construir mais usinas de energia e ampliar as opções disponíveis. Após uma década de demanda energética estável nos Estados Unidos, uma análise do Goldman Sachs descreveu a nova onda de data centers como responsável por "um tipo de crescimento de consumo de eletricidade que não se via fazia uma geração".[23] As empresas de serviços públicos estão adiando a desativação de usinas a gás e carvão, assim como a transição para energia renovável.[24] A Microsoft

reativou a Three Mile Island, uma usina nuclear perto de Middletown, Pensilvânia, que sofreu um acidente nuclear no final da década de 1970, o pior acidente nuclear comercial da história do país. Até 2030, no ritmo atual de crescimento, estima-se que os data centers usarão 8% da energia dos Estados Unidos, em comparação com 3% em 2022;[25] mundialmente, a computação de IA poderá consumir mais energia do que toda a Índia, o terceiro maior consumidor de eletricidade do mundo.

Essa escala — a mega-hiperescalabilidade — gerou consequências ambientais surpreendentes. No entanto, nesse mesmo momento, o encobrimento desse impacto por parte das corporações atingiu novos patamares. Desde o artigo de Emma Strubell e a citação de Gebru em "Papagaios Aleatórios", os gigantes da tecnologia passaram a ocultar ainda mais os detalhes técnicos de seus modelos, tornando difícil estimar e rastrear suas pegadas de carbono. Ao mesmo tempo, essas empresas intensificaram campanhas direcionadas ao público e a formuladores de políticas, apoiando-se em contra-argumentos poderosos: os data centers se tornarão tão eficientes que o impacto deixará de ser um problema; a IA generativa abrirá caminho para novas inovações climáticas; a IAG resolverá as mudanças climáticas de uma vez por todas.[26]

Embora a última afirmação seja impossível de provar,[27] as duas primeiras são altamente enganosas, afirma Sasha Luccioni, pesquisadora científica e líder climática na empresa de IA de código aberto Hugging Face. A segunda é especialmente perniciosa: de fato, como catalogado pela iniciativa que depois se tornou a organização sem fins lucrativos Climate Change AI, existem muitas tecnologias de IA[28] capazes de acelerar a sustentabilidade, mas raramente são tecnologias de IA *generativas*. "O que você precisa para o clima são modelos de aprendizado supervisionado, ou modelos de detecção de anomalias, ou até modelos estatísticos de séries temporais", diz Luccioni, que também é fundadora da iniciativa. Todos esses modelos pertencem às gerações anteriores das tecnologias de IA, principalmente ferramentas de aprendizagem de máquina, que são pequenas, energeticamente eficientes e, em alguns casos, podem até ser executadas em um laptop potente. "A IA generativa tem uma pegada energética e de carbono extremamente desproporcional, com pouquíssimos aspectos positivos para o meio ambiente", acrescenta ela.

Luccioni afirma que seus antigos colaboradores em empresas de capital fechado já não recebem mais permissão de seus empregadores para assinar artigos como coautores dela sobre o impacto ambiental da IA. Em vez disso, ela tem trabalhado com colaboradores externos e acadêmicos como Strubell, cuja pesquisa foi sua inspiração inicial. Em um desses artigos,[29] escrito juntamente com Yacine Jernite, líder de aprendizagem de máquina e sociedade na Hugging Face, eles mediram a pegada de carbono da execução de modelos de IA generativa de código aberto como um proxy para o que essas empresas estão desenvolvendo. Descobriram que a produção de mil textos gerados por esses modelos consumia, em média, energia suficiente para carregar quase

quatro vezes um smartphone comum.[30] A geração de mil imagens consumia, em média, energia equivalente a 242 cargas completas de smartphone; em outras palavras, cada imagem gerada por IA poderia usar energia suficiente para carregar cerca de 25% de um smartphone.[31] Os artigos de Luccioni e Strubell estão entre os poucos que ainda oferecem medidas quantificáveis da emissão de carbono associada aos modelos de IA generativa. É uma batalha constante. Em determinado momento, ela entrou em contato com mais de quinhentos autores de artigos recentes sobre aprendizagem de máquina para solicitar informações básicas sobre o treinamento de seus modelos. "Quase ninguém respondeu", conta ela. "As pessoas simplesmente não respondiam ou diziam que essa era uma informação confidencial."

Mesmo que as empresas de hiperescalabilidade falem abertamente sobre a sustentabilidade de sua infraestrutura de computação,[32] executivos da Microsoft admitem internamente que a disponibilidade intermitente de energia renovável não é suficiente quando os data centers precisam operar 24 horas por dia, 7 dias por semana. A operação contínua é considerada tão crucial que Google, Amazon e, mais recentemente, a Microsoft passaram a construir seus campi em trios, garantindo um backup para o backup caso alguma instalação apresente falhas.[33]

Durante a passagem do furacão Irma, na Flórida, e do furacão Harvey, no Texas, mesmo com milhões de pessoas sem eletricidade, hospitais evacuando pacientes e centenas de milhares de residências e empresas lidando com danos e destruição, os data centers nessas regiões continuaram funcionando.[34] Funcionaram tão bem que familiares desabrigados de um funcionário da instalação ficaram ali durante o desastre natural.

A terra e a energia necessárias para sustentar esses megacampi são apenas dois dos insumos da cadeia de suprimentos global que sustenta a expansão dos data centers. O mesmo vale para o volume extraordinário de minerais, incluindo cobre e lítio, exigidos para construir o hardware: computadores, cabos, linhas de energia, baterias e geradores de backup, e também para o volume igualmente extraordinário de água doce — sim, potável — necessária para resfriar os servidores e gerar eletricidade para que funcionem. A água usada para a refrigeração geralmente deve ser limpa o suficiente para evitar o entupimento de tubulações e o crescimento de bactérias; a água potável atende a esse padrão. De acordo com uma estimativa de pesquisadores da Universidade da Califórnia,[35] em Riverside, o aumento da demanda por IA pode usar, até 2027, entre 6 e 3 trilhões de litros de água predominantemente doce por ano no mundo, o equivalente a metade da água usada anualmente pelo Reino Unido. Esses efeitos não serão distribuídos de forma uniforme. Outro estudo constatou que, nos Estados Unidos, um quinto dos data centers já retirava essa água, antes mesmo do boom da IA generativa, de bacias hidrográficas moderada ou altamente estressadas por seca ou outros fatores.[36] Em países do Sul, como o Chile, muitas vezes são as comunidades mais vulneráveis que vêm sofrendo o impacto direto dessas economias de extração

aceleradas.

À medida que mais comunidades começaram a perceber o impacto dos data centers em suas vidas, um número crescente delas se posicionou contra o desenvolvimento irrestrito. Em resposta,[37] os desenvolvedores desses centros passaram a adotar táticas mais sofisticadas para manter tudo funcionando como sempre: entraram nas comunidades de forma discreta, usando empresas de fachada; fizeram doações para programas locais a fim de reduzir a resistência; e apresentaram promessas sobre a sustentabilidade de suas operações, promessas que, uma a uma, foram sendo abandonadas depois que os projetos avançaram e se tornaram difíceis de reverter.

Em um caso na Virgínia,[38] um grupo de moradores que protestava contra diversos data centers gigantes ficou chocado ao descobrir um e-mail enviado pelo advogado de um dos desenvolvedores sugerindo colocar os manifestantes sob vigilância. "Precisamos de um ou dois informantes nesse grupo",[39] escreveu o advogado.

A partir do momento em que decidiu apostar na expansão, a OpenAI passou a buscar uma quantidade sem precedentes de infraestrutura de computação. "Na IA, quem tiver o maior computador obtém os maiores benefícios",[40] disse Brockman em 2019. Assim, em parceria com a Microsoft, Altman desenvolveu um plano para que a gigante da tecnologia suprisse a demanda exponencialmente crescente da OpenAI por poder computacional. As duas empresas trabalhariam juntas no projeto e no fornecimento de dezenas de supercomputadores de pesquisa — data centers que precisariam ser equipados com chips Nvidia para treinar diversos modelos de IA ao longo das investigações da OpenAI. E, de forma crucial, a Microsoft também construiria uma série de supercomputadores cada vez maiores e mais potentes para treinar cada nova geração dos modelos da OpenAI.

Altman passou a se referir à série de supercomputadores como "fases", chegando a mostrar aos funcionários um slide que ilustrava o tamanho de cada uma delas, com a Fase 5, a última, incrivelmente maior do que todas as anteriores.[41] O supercomputador usado pela OpenAI para treinar o GPT-3 correspondia à Fase 1. Equipado com dez mil V100s,[42] foi construído em West Des Moines, Iowa, onde a Microsoft havia se instalado pela primeira vez em 2012. A empresa causou boa impressão nas autoridades locais ao aportar uma quantia considerada "impressionante" em pagamentos de impostos, segundo o então prefeito, o que ajudou a cidade a realizar grandes melhorias em sua infraestrutura pública. Ao longo de mais de uma década, a Microsoft também investiu cerca de 2,5 milhões de dólares em programas comunitários locais, em sua maioria de organizações sem fins lucrativos.[43] O supercomputador da Fase 1 foi nomeado como código Odyssey pela Microsoft; a OpenAI o chamou de Owl (coruja), seguindo uma convenção adotada desde o início de nomear em ordem alfabética cada um dos clusters de computação da divisão de Pesquisa com nomes de animais. Quando as 26

letras finalmente se esgotaram, a equipe passou a batizá-los com nomes de elementos da tabela periódica, ordenados pelo número atômico.

A Fase 2 também foi construída em Iowa e usada para treinar o GPT-4. Ela começou com os dezoito mil A100s mencionados no roteiro de pesquisa de 2021 e acabou crescendo para cerca de 25 mil até o fim do processo de treinamento do modelo. Na Microsoft, a Fase 2 recebeu o nome Telêmaco, em referência ao filho de Odisseu na mitologia grega. Na OpenAI, foi nomeada como Raven (corvo).

A Fase 3 foi transferida para o Arizona. Com terrenos baratos, incentivos fiscais e proximidade da Califórnia, o estado rapidamente se tornou um dos principais pontos de expansão para data centers entre todos os provedores de serviços de nuvem. Depois de cultivar cuidadosamente relações favoráveis com os governos de duas cidades subdesenvolvidas nos arredores de Phoenix,[44] a Microsoft comprou três terrenos entre 2018 e 2019 e fez doações semelhantes às de Iowa para várias organizações comunitárias, a fim de amenizar possíveis objeções dos moradores. Os três terrenos somavam quase 600 acres, o equivalente a mais de 450 campos de futebol. (A Microsoft ampliaria essa área com mais 283 acres em 2024.) Cada terreno abrigaria um novo campus de data center destinado a atender clientes do Azure, além de dar suporte à OpenAI na região de nuvem conhecida como "West US 3".

A Fase 3 — chamada de Inglewood pela Microsoft e Whale pela OpenAI — custaria vários bilhões de dólares para ser construída e abrigaria centenas de milhares de Nvidia H100s, a geração seguinte aos A100s, destinados a treinar o que a OpenAI acreditava, à época, que seria lançado como GPT-4.5 e GPT-5.

Para a Fase 4, planejada em Wisconsin, Altman estimava custos na casa dos 10 bilhões de dólares, usando os então recentes chips Nvidia B100, uma geração além dos H100. Era uma quantia impressionante, considerando que os data centers de hiperescala mais caros do período custavam entre 1 bilhão e 2 bilhões de dólares. Altman não pretendia parar por aí, chegando a mencionar, quase de passagem, a ideia de um supercomputador de 100 bilhões de dólares para a Fase 5. Após o enorme sucesso do ChatGPT, porém, ele moderou as expectativas. Com tantos chips reservados para atender aos clientes do ChatGPT, a Microsoft estava tendo dificuldade para adquirir GPUs com rapidez suficiente para acompanhar o ritmo de construção da Fase 3. O Whale acabou sendo dividido em três clusters separados — Beluga, Narwhal e Orca —, com planos de concluir a construção em algum momento de 2024.

Ninguém dentro da Microsoft ou da OpenAI sabia se a Fase 5 era tecnicamente possível. Nos planos de projeto das duas empresas,[45] segundo reportagem posterior da *The Information*, a instalação de 10 bilhões de dólares poderia exigir até 5 mil megawatts, quase igualando a demanda média de energia de toda a cidade de Nova York. Embora o plano não parecesse o mais sólido financeiramente como investimento comercial, o principal gargalo não

era o dinheiro. Era a energia. “Estamos ficando sem terra e sem energia”, diz um funcionário da OpenAI. Dentro de ambas as organizações, entendia-se que a Fase 5 só seria viável com algum tipo de inovação. Ou a Microsoft e a OpenAI precisariam dividir o supercomputador em vários campi para distribuir a demanda energética e descobrir como treinar um modelo de IA em locais distantes, ou, como Altman às vezes gostava de dizer, o problema se resolveria por si só com um futuro avanço em fusão nuclear.

Às vezes, Altman compartilhava com os funcionários da OpenAI notícias otimistas sobre a Helion Energy, a startup de fusão nuclear que representava seu maior investimento pessoal, e da qual a Microsoft já havia se comprometido a comprar energia assim que a usina, com meta de geração de 50 megawatts, entrasse em operação. O *Wall Street Journal* relataria mais tarde que a OpenAI e a Helion também estavam em negociações para fechar um acordo, do qual Altman havia se retirado.[46]

Altman e outros executivos nunca mencionaram o impacto ambiental dos data centers nas reuniões gerais da empresa. Enquanto a OpenAI treinava o GPT-4 em Iowa, o estado já enfrentava dois anos de seca. A Associated Press relatou posteriormente que, durante um único mês de treinamento do modelo, os data centers da Microsoft consumiram cerca de 43 milhões de litros de água, aproximadamente 6% de toda a água do distrito.[47] O GPT-4 permaneceu em treinamento no local por três meses. (Um porta-voz da Microsoft afirmou que a empresa está trabalhando para aumentar a eficiência hídrica em 40% acima da linha de base de 2022 e que reporia mais água do que consome em suas operações globais até 2030, com foco especial nas regiões onde atua e que enfrentam escassez hídrica.)[48]

O Arizona também enfrenta uma grave crise hídrica. Em 2022, enquanto a Microsoft preparava o terreno para a Fase 3,[49] um estudo publicado na revista *Nature Climate Change* revelou que o sudoeste dos Estados Unidos vivia a pior seca dos últimos mil anos. Essa situação, combinada a uma gestão profundamente inadequada, levou o rio Colorado, do qual o Arizona e outros seis estados dependem para obter água potável, a níveis perigosamente baixos. Sem medidas drásticas, o rio pode deixar de fluir.[50] A escassez de água agrava a crise energética, já que as mudanças climáticas vêm atingindo a região com ondas de calor recordes e implacáveis, levando as famílias a aumentarem o uso de ar-condicionado. A região depende, em parte, da energia hidrelétrica da represa Hoover e de usinas nucleares resfriadas a água; ou seja, precisa de água justamente para produzir mais energia. Em 2023, a área metropolitana de Phoenix quebrou diversos recordes de calor e registrou o pior ano no que se refere a mortes relacionadas às altas temperaturas, que aumentaram pelo menos 30% em relação a 2022, ultrapassando as seiscentas.[51] “Todas essas coisas”, diz Tom Buschatzke, diretor do Departamento de Recursos Hídricos do Arizona, “estão convergindo em uma direção desafiadora.”[52]

O que Altman mencionava, na verdade, era sua impaciência. Em março de 2024, depois de ter passado os primeiros anos da corrida pela IA generativa

praticamente adormecida, a Meta surgiu com planos agressivos para colocar 350 mil H100s em operação como parte de uma expansão de infraestrutura ainda maior destinada a sustentar sua súbita explosão de investimentos em IA generativa.[53] Isso representava um número maior de GPUs do que a OpenAI tinha disponível. Altman não ficou satisfeito. Para ele, a Microsoft estava avançando devagar demais, e essa lentidão estava custando à OpenAI a vantagem competitiva.

Quando Sonia Ramos era criança, testemunhou um acidente que moldaria toda a sua vida. Nascida em uma família de mineiros no Chile,[54] o pai trabalhava para uma mineradora de cobre dos Estados Unidos, e ela cresceu cercada pelos filhos de outros trabalhadores. Em 1957,[55] parte da mina de Chuquicamata desabou, matando várias pessoas e ferindo dezenas. Embora o pai tenha sido poupado, ela se lembra de ver a miséria das consequências se espalhar ao seu redor: famílias afetadas caindo na pobreza extrema, crianças definhando de fome.

Quatro décadas depois, quando Ramos começou a protestar contra a mineração, tornando-se uma das vozes indígenas mais ativas e contundentes do Chile ao denunciar a destruição social, cultural e ambiental do país, ela voltou às lições aprendidas naquela tragédia: a indústria mineradora é um sistema e, deixado à própria sorte, esse sistema buscará lucro a qualquer custo. "O trabalhador não existe", diz ela. Nenhuma das vítimas recebeu qualquer homenagem ou cerimônia; nenhuma família recebeu indenização. "Naquele lugar, não há humanidade."

O Chile é o maior produtor de cobre do mundo, responsável por um quarto do abastecimento global. Desde o início, a mineração desse minério transformou não apenas a terra, mas também as sociedades que dela dependem. Às vezes, os efeitos são visíveis: Chuquicamata é hoje a maior mina de cobre a céu aberto do planeta, uma ferida aberta na terra que os explosivos aprofundam continuamente. As rochas removidas,[56] empilhadas em montanhas imponentes que monumentalizam as cavidades de onde foram extraídas, estão lentamente soterrando os restos de uma cidade que foi abandonada depois que a expansão da mineração começou a engoli-la. A atividade também drenou a água da região para processar o cobre.[57] Em certo momento, uma empresa multinacional estrangeira consumiu tanta água que esgotou toda uma bacia em um salar próximo, destruindo o rico ecossistema.

Menos visíveis são os rastros de arsênico que a indústria deixa no ar e na água,[58] algo que elevou as taxas de câncer em todo o norte do país, e as maneiras pelas quais a mineração reestruturou a vida indígena e semeou divisões entre diferentes comunidades. Com suas terras esgotadas de água e minerais, os atacamenhos, designação que reúne os diversos grupos indígenas que compartilham essa região, já não conseguem se sustentar cultivando as próprias plantações ou criando o próprio gado. Essa transformação mergulhou

suas cidades em profunda pobreza.[59] O crime aumentou, assim como a depressão, o alcoolismo e a delinquência. Falta comida, falta água encanada, faltam cuidados de saúde adequados e recursos educacionais, tendo em vista os benefícios quase nulos dos bilhões de dólares em lucros que suas terras geraram para outros. Em vez disso, muitos são forçados a trabalhar para a mesma indústria que tomou seus territórios e a receber atendimento nas pequenas clínicas que ela própria financia.[60] Onde antes havia maior unidade, os grupos indígenas disputam recursos cada vez mais escassos.

O lítio é uma descoberta mais recente na região,[61] encontrado por acaso por uma empresa dos Estados Unidos na década de 1960, enquanto buscava a água necessária para a mineração de cobre. Ao perfurar os salares, a companhia encontrou altas concentrações de lítio flutuando em uma salmoura oleosa abaixo da superfície, abrindo uma nova frente de extração e acelerando o esgotamento de ainda mais ecossistemas. Hoje, o Chile produz cerca de um terço do lítio mundial, ficando atrás apenas da Austrália.[62] O material é extraído principalmente do salar de Atacama,[63] a maior planície salgada do país, por meio do bombeamento de sua salmoura para piscinas cintilantes de cor turquesa para que o sol evapore a água e cristalize a solução em lítio e outros subprodutos. Os salares já foram lar de bandos de flamingos, que os atacamenhos consideram irmãos espirituais. Contudo, eles desapareceram.[64] A jovem filha de um líder indígena da comunidade de Peine tem apenas as histórias dos antepassados e um flamingo de pelúcia para lembrar que um dia eles estiveram ali.

Ao longo dos anos, os atacamenhos ouviram inúmeras narrativas usadas para justificar toda essa extração. Em 2022, quando a União Europeia estabeleceu novas políticas para a transição energética e a demanda por lítio disparou,[65] empresas e políticos, no Chile e no resto do mundo, passaram a exaltar a importância da indústria mineradora do país para impulsionar um futuro melhor. As comunidades indígenas, porém, vendo suas terras e seus modos de vida serem destruídos, perguntavam: um futuro melhor para quem? "A população local nunca tem a capacidade de pensar no próprio destino fora das forças da economia e da política internacional",[66] diz Cristina Dorador, microbiologista que vive no norte do Chile e estuda sua rica biodiversidade.

Hoje, as mesmas narrativas estão sendo recicladas com a IA generativa. A extração acelerada de cobre e lítio para construir os megacampi — e para erguer usinas de energia e milhares de quilômetros adicionais de linhas de transmissão que os sustentem — está, segundo o Vale do Silício, inaugurando um futuro melhor e mais brilhante.[67] Bloquear essa extração seria, portanto, bloquear o progresso fundamental da humanidade. Contudo, não é a mineração em si que as comunidades indígenas rejeitam. "Nossos ancestrais eram mineradores", diz Ramos. Foram eles que descobriram o cobre. O problema, afirma ela, é a escala.

Essa escala consumiu tudo. Tornou o norte e o restante do Chile completamente dependentes da indústria, impedindo o surgimento de outras

economias. Ela sufocou a capacidade do país e do mundo de imaginar caminhos diferentes, nos quais o desenvolvimento pudesse existir sem a pilhagem dos recursos naturais, afirma Ramos. Ao permitir a produção de modelos massivos de IA generativa, essa mesma escala também perpetuou estereótipos racistas sobre povos indígenas que já sofrem com a maneira como a tecnologia é construída. No Brasil, uma exposição de arte de 2023,[68] coproduzida por uma universidade chilena, evidenciou o vasto abismo entre a realidade das ricas culturas indígenas da América Latina e as representações lamentavelmente distorcidas geradas por ferramentas como Midjourney e Stable Diffusion, imagens que retratavam esses povos como primitivos e tecnologicamente ineptos.

Nos últimos anos, os atacamenhos têm demonstrado uma crescente resistência. Eles hasteiam bandeiras pretas em suas casas para denunciar a exploração das terras e da comunidade. Organizaram protestos para bloquear as estradas pelas quais passam os ônibus e caminhões das empresas rumo às minas. Contrataram advogados para fazer valer seus direitos legais como povos indígenas segundo o direito internacional, que protege a soberania cultural e territorial. À medida que as empresas e o governo chileno foram obrigados a incluí-los nas negociações, o ponto central das demandas indígenas passou a ser a necessidade de o governo realizar pesquisas sobre a saúde dos ecossistemas do deserto do Atacama, quantificando a perda de água e quaisquer danos irreparáveis.[69]

Ramos também criou a própria fundação, que reúne "o ancestral e o não ancestral", como ela mesma diz, para promover e conduzir pesquisas científicas sobre a riqueza natural que o deserto do Atacama tem a oferecer. Devido às condições extremas e únicas, a região abriga numerosas comunidades microbianas — potencialmente úteis para a obtenção de medicamentos ou novas fontes de energia — que não existem em nenhum outro lugar. Pelas mesmas razões, o deserto também vem sendo estudado há décadas como um análogo climático de Marte. Ela espera que essas descobertas ajudem a demonstrar o valor de preservar sua bela terra natal. Contra as narrativas de progresso acelerado que têm servido para justificar a extração, ela propõe novas concepções de progresso, voltadas à cura, à sustentabilidade e à regeneração.

Enquanto a luta de Ramos continua no norte, uma batalha diferente está sendo travada no coração do Chile: a disputa em torno da aceitação pelo governo dos data centers da indústria de tecnologia. Quanto mais rapidamente os hiperescaladores se expandiram, ultrapassando a disponibilidade de terra e energia em suas regiões tradicionais de operação, mais agressivamente passaram a reivindicar esses recursos em novos territórios ao redor do mundo.

A Microsoft, sozinha, gastou mais de 55 bilhões de dólares no ano fiscal de 2024, quase um quarto de sua receita declarada, para construir o que a *SemiAnalysis* descreveu como "a maior construção de infraestrutura que a humanidade já viu".[70] O Google, por sua vez, afirmou na teleconferência de

resultados do terceiro trimestre de 2024 que pretendia elevar os gastos com data centers para cerca de 50 bilhões de dólares naquele ano fiscal.[71] A Meta afirmou que provavelmente encerraria o ano com até 40 bilhões de dólares investidos em expansões de data centers e infraestrutura, valor que, segundo suas estimativas, aumentaria no ano seguinte.[72]

Em uma rara tarde nublada de junho de 2024, Alexandra Arancibia dirige o carro por Quilicura, um município nos arredores de Santiago onde ela mora e atua como vereadora, até aquilo que considera o símbolo definitivo da enorme disparidade de poder entre as gigantes americanas da tecnologia e sua comunidade. A menos de trinta minutos dos bairros mais pitorescos de Santiago,[73] repletos de cafés em estilo europeu e restaurantes veganos, as estradas à nossa frente são horríveis e não têm manutenção, com montanhas de lixo espalhadas ao longo de lixões ilegais controlados por uma máfia local.

Depois de passarmos por um cemitério dedicado a animais de estimação falecidos, chegamos ao que parece ser um terreno abandonado, com tufos de arbustos e algumas árvores esparsas, carentes de nutrientes, brotando do solo. Na maior parte dos dias, a terra é tão seca que lembra o deserto do Atacama; neste dia, a chuva transforma tudo em lama.

No meio do terreno, uma placa roxa anuncia em espanhol: "Bem-vindo à Floresta Urbana de Quilicura", um projeto que, como explicado, o Google iniciou em 2019 como retribuição à comunidade por hospedar seu data center. A placa inclui um diagrama destinado a explicar os benefícios da "floresta": no lado esquerdo, uma ilustração mostra Quilicura como uma zona industrial repleta de fábricas que produzem gases de efeito estufa e poluição do ar; à direita, a floresta aparece florescendo sob a chuva generosa que cai de uma grande nuvem chamada "SMOG".

O Google se orgulha dessa floresta em seu site e em seus comunicados de relações públicas. Quando peço à porta-voz da empresa no país uma entrevista sobre o desenvolvimento do Google no Chile,[74] ela me envia alguns materiais informativos bem produzidos, acrescentando depois que a empresa cria um programa de impacto comunitário para cada novo data center, com o objetivo de apoiar projetos locais nas áreas de educação, sustentabilidade, acesso à internet e saúde; no caso de Quilicura, o Google investiu mais de 1,2 milhão de dólares. Nos materiais, a parte dedicada à floresta fala sobre os moradores que utilizam o espaço verde. Não há moradores. O local fica muito distante de qualquer linha de ônibus, e não há casas suficientes na área circundante para que sejam mencionadas. Fora do modesto terreno, pequeno demais até para abrigar o próprio data center do Google, uma dúzia de cães vagueia, latindo e revirando o lixo. A porta-voz afirmou que a floresta está sendo atualizada para "aprimorar a experiência para a comunidade".

Arancibia sorri com ironia enquanto observamos a cena. "Você se sente no Vale do Silício?"

Arancibia tinha acabado de entrar na faculdade quando percebeu que

Quilicura era um lugar onde as coisas eram descartadas. Todos os dias, no trajeto até a universidade, ela passava por partes do município repletas de lixo que nem sabia que existiam. Ela nunca havia pensado em Quilicura como "lar"; era apenas o lugar onde vivia, uma cidade operária, subdesenvolvida e sem graça, para onde os pais tinham se mudado quando ela era pequena. No entanto, perceber que o local era tratado como um verdadeiro lixão despertou nela um profundo desejo de revitalizar a terra e devolver-lhe sua antiga beleza.

Há apenas duas décadas,[75] quando Arancibia era criança, Quilicura era majoritariamente rural: pastagens bucólicas e pântanos cintilantes, abrigo de uma biodiversidade distinta, mas tão rica quanto a do deserto do Atacama — aves, animais e diversas variedades de flora. Então, chegou a máfia do lixo, permitindo que qualquer pessoa de Santiago despejasse seus resíduos em Quilicura por um preço. Alguns aterros operam há tanto tempo que grama e ervas daninhas cresceram sobre eles, fazendo-os parecerem colinas deformadas e inquietantes que avançam pela paisagem. Depois vieram diferentes indústrias, incluindo cervejarias e incorporadoras imobiliárias, que desviaram ainda mais terras e extraíram água dos pântanos para os próprios fins. Hoje, apenas pequenas áreas verdes espalhadas por esse município de 57 quilômetros quadrados oferecem um vislumbre do que Quilicura já foi um dia.

Nesse contexto, o Google chegou a Quilicura em 2012 para construir seu primeiro data center na América Latina. Em seu site, a empresa apresenta com orgulho o data center,[76] que entrou em operação em janeiro de 2015, como um dos mais eficientes e ecológicos do continente. Na época, ninguém na região prestou muita atenção ao projeto; e, certamente, ninguém no epicentro mais rico de Santiago estava atento ao que acontecia ali. Os moradores que passavam pelo data center em seus trajetos de ônibus para o trabalho presumiam que se tratava de uma fábrica que produzia cerveja ou alimentos e que ofereceria empregos à comunidade local.

O data center do Google — como a maioria dos data centers — não gerou muitos empregos além dos postos criados durante a construção inicial. Uma vaga de 2024 para técnico mecânico, um dos poucos cargos de longo prazo disponíveis, foi anunciada no mural de empregos do Google apenas em inglês; o anúncio afirmava que a companhia não consideraria currículos enviados em qualquer outro idioma. O local, como apontam ativistas, também não trouxe muitos outros benefícios para a comunidade local. Nas proximidades, as escolas públicas ainda carecem de internet de qualidade ou de dispositivos que permitam que os alunos acessem a rede.

A chegada do data center marcou Quilicura e o restante de Santiago como um destino desejável para a expansão física do Vale do Silício. Em 2019, o Google anunciou que construiria seu segundo data center latino-americano na área metropolitana da capital. Pouco depois, Microsoft e Amazon anunciaram que também se instalariam ali. O governo chileno foi rápido em recebê-las, posicionando o país como um refúgio seguro e estável para investimento estrangeiro direto na América Latina, região que sofre com a reputação de

governos pouco confiáveis e instabilidade social e econômica. Em 2020, o governo deu um passo além: anunciou um projeto para construir um novo cabo submarino,[77] uma espécie de rodovia de dados, para conectar Ásia e Pacífico diretamente às Américas pela costa central do Chile, não muito longe de Santiago. O país se tornaria um centro global de infraestrutura digital. O Google apoiou a parceria.[78]

No entanto, em julho de 2019, quando o Google iniciou a documentação para seu segundo data center no Chile, um grupo de moradores estava atento. A empresa havia escolhido Cerrillos como o novo local, outro município da classe trabalhadora na fronteira com Santiago. Igual a Quilicura, Cerrillos tem uma longa história de desatenção e abandono. Entre as décadas de 1930 e 1990,[79] uma fábrica de cimento pertencente a uma empresa belga contaminou a comunidade com níveis letais de amianto, levando ao que um historiador chileno chamou de "o maior genocídio industrial" do país. Ainda hoje, os moradores morrem em decorrência de câncer em taxas acima da média. Além disso, Cerrillos é especial: em um país onde a água é privatizada, o município abriga o único serviço público do Chile, que fornece a água subterrânea local para comunidades vizinhas e, em situações de emergência, para outras regiões do país. Essa combinação única — uma história de negligência e uma fonte de água preciosa — criou um terreno fértil para o surgimento de diversos grupos ativistas ambientais, acostumados à vigilância constante e ferozmente protetores dos recursos diante de qualquer forma de extração. Naquele verão, quando o Google apresentou um relatório à agência ambiental do Chile para obter aprovação para o data center — um processo em grande parte burocrático —, o Mosacat, um grupo ativista dedicado à defesa da água, começou a analisar minuciosamente todas as 347 páginas do documento.[80] Escondida em meio ao relatório, estava a informação de que o data center planejava usar cerca de 169 litros de água potável por segundo para resfriar os servidores. Em outras palavras, a instalação poderia consumir mais água do que toda a população de Cerrillos, cerca de 88 mil residentes, ao longo de um ano.[81] Para o Mosacat, isso era inaceitável. A instalação não apenas retiraria essa água diretamente da fonte pública de Cerrillos, como faria isso num momento em que todo o abastecimento de água potável do país estava ameaçado. Em 2019, assim como Iowa e Arizona, o Chile já enfrentava uma megasseca devastadora e historicamente sem precedentes, que perdurava havia nove anos.[82]

Tania Rodríguez, integrante do Mosacat, me entrega todas as 347 páginas do relatório ambiental do Google, impressas e encadernadas entre duas capas plásticas azuis. O volume cai no meu colo com um baque, uma manifestação física da forma como o Vale do Silício usa o conhecimento técnico para justificar a tomada de decisões centralizada. Da parte inferior do bloco, vários post-its cuidadosamente rotulados se projetam. "Agua potable" [água potável], diz um deles, marcando as páginas que tratam da necessidade do data center

de consumir água doce para fins de refrigeração.

O Mosacat foi fundado em 2019,[83] depois que ativistas de vários movimentos, em defesa dos direitos das mulheres, da moradia, dos trabalhadores e do meio ambiente, se uniram em solidariedade para formar um coletivo unificado. Muitos deles haviam se conhecido durante um protesto contra um projeto de mineração ilegal. Segundo o grupo, a atuação do Mosacat conseguiu expulsar os mineradores, encerrar o projeto e transformar a área em uma reserva natural protegida. Pouco tempo depois, um amigo do coletivo, que hoje é membro do Congresso Nacional do Chile, informou-os sobre o projeto do data center do Google e os incentivou a analisar o consumo de água previsto.

Os integrantes do Mosacat não são tecnólogos. Ainda assim, eles leram todas as páginas repletas de diagramas densos e terminologia técnica misteriosa, fizeram anotações detalhadas e memorizaram os meandros dos data centers e dos sistemas de refrigeração, a fim de se prepararem para enfrentar o Google. Rodríguez dá uma risada animada quando pergunto como conseguiram digerir tantas informações. "Todos nós fomos necessários", diz ela, referindo-se aos mais de uma dúzia de voluntários que compõem o Mosacat e realizam esse trabalho nas horas roubadas entre empregos e obrigações familiares.

Na maior parte dos casos, projetos no Chile que demandam uso de água levam muito tempo para obter aprovação. No caso do Google, porém, a autorização veio rapidamente, mesmo em meio a uma série de emergências hídricas provocadas pela seca. No início, o Mosacat tentou contestar o projeto por meio do parceiro local do Google, uma empresa chilena de investimentos e serviços chamada Dataluna. A primeira reunião foi ruim, relata o grupo: os representantes da Dataluna pareciam ter pouco conhecimento sobre o projeto e negaram que ele utilizaria água doce.

A partir daí, o Mosacat levou a questão ao governo local. O prefeito e a Câmara Municipal já haviam se reunido com a Dataluna, lembra o grupo, e tinham a falsa impressão de que o data center precisaria apenas de águas residuais para seu resfriamento. Após serem informados pelo Mosacat, os representantes do governo, alarmados, exigiram esclarecimentos da Dataluna. O assunto então escalou: saiu das mãos da Dataluna, passou para a divisão chilena do Google e acabou chegando até a sede da empresa em Mountain View, na Califórnia.

Em outubro de 2019, o Google enviou dois engenheiros e um advogado a Cerrillos para fazer uma apresentação à comunidade. No dia da visita, o Mosacat cobriu as ruas com cartazes de protesto ao longo do trajeto que os funcionários do Google percorreriam até o local da reunião. No encontro em si, o Mosacat não estava sozinho. Entre as cerca de duas dúzias de moradores presentes, havia representantes de outros seis grupos ativistas e comunitários. Os engenheiros do Google eram gringos, altos e falavam apenas inglês, recorda o Mosacat. Eles fizeram uma apresentação altamente técnica,

enquanto o advogado da empresa também atuava como tradutor. Durante as discussões, os membros do Mosacat que falavam inglês afirmam ter ouvido o advogado traduzindo mal suas falas. Em certo momento, os representantes do Google tentaram acalmar a comunidade oferecendo-se para plantar uma floresta urbana semelhante à que a empresa havia instalado em Quilicura. O gesto não impressionou o Mosacat nem os demais grupos: o Google não parecia estar ali para se envolver de fato ou ouvir o que a comunidade desejava. "Eles vieram para nos intimidar", diz Alejandra Salinas, integrante do Mosacat e vereadora do município vizinho a Cerrillos, Maipú. "Pense nisso. Eles vêm nos oferecer árvores enquanto secam nossa terra."

Os moradores de Cerrillos não precisavam de árvores. Não precisavam que o Google construísse um parque para eles — como se a área metropolitana de Santiago fosse um lugar tão atrasado que ainda não tivesse parques. Eles precisavam que o Google parasse de tratar as terras como um local a ser saqueado para a obtenção de água preciosa e outros recursos; precisavam que a empresa deixasse de descartar a comunidade como mera espectadora, em vez de participante no desenvolvimento dos próprios projetos locais. "Sabemos que alimentamos o mundo, que fornecemos matérias-primas — por exemplo, cobre e lítio", pondera Salinas. "Ninguém está dizendo que nosso tesouro é só nosso e que não vamos compartilhá-lo. Sim, podemos ajudar uns aos outros. Mas eles não vão vir aqui, usar a água, que é vital para a vida, e nos deixar sem nada."

Na época, o Chile atravessava uma forte turbulência política que duraria meses, conhecida como *Estallido Social*. Protestos explosivos e, por vezes, violentos ocorriam semanalmente, começando naquele mesmo outubro de 2019: um clamor coletivo contra o desemprego, a privatização e a profunda desigualdade, que deixou milhares de feridos e dezenas de mortos. No fim de 2019, à medida que comunidades em todo o país começaram a realizar referendos em resposta ao movimento para reformar a política local e nacional, o Mosacat viu uma oportunidade. Aproveitou o referendo em Cerrillos para acrescentar uma pergunta sobre se os moradores concordavam que o Google construísse uma instalação que consumiria tamanha quantidade de água da comunidade. Eles se mobilizaram para divulgar a verdadeira natureza do projeto, distribuindo panfletos nas esquinas, batendo de porta em porta e espalhando cartazes por todo o município.

Em dezembro desse mesmo ano, o Mosacat venceu: o referendo rejeitou a construção do data center por uma pequena maioria de votos. Em 2020, o governo uniu-se ao Mosacat em uma ação judicial contra o projeto em um tribunal ambiental.

Na reunião com o Google, seus representantes acabaram demonstrando uma atitude mais amigável, afirma o Mosacat, mostrando maior disposição para negociar assim que perceberam que alguns moradores compreendiam o que estava sendo discutido. Contudo, como me disseram Julian Posada, professor assistente em Yale, e Mercy Mutemi, advogada de Mophat Okinyi,

o risco para um país do Sul ao resistir é sempre o mesmo: uma empresa do Vale do Silício pode simplesmente pegar seu dinheiro e ir embora. À medida que o projeto permanecia paralisado e a demanda do Google por mais computação se intensificava, a empresa anunciou que transferiria para outro país o próximo data center planejado.

No Uruguai, um pequeno país com 3,4 milhões de habitantes, a empresa nacional de telecomunicações Antel possui três data centers que ajudam a fornecer serviços de internet e telefonia celular para toda a população. Somados, eles ocupam cerca de 5 mil metros quadrados. Um deles, com menos de mil metros quadrados, está situado em uma rua residencial típica de Montevidéu, mais alto e mais largo que os edifícios vizinhos, mas ainda assim integrado ao entorno.[84]

O data center funciona com duas linhas de energia separadas das demais da vizinhança. Trinta por cento do espaço são ocupados por computadores; os outros 70% abrigam escritórios administrativos, closets elétricos e salas mecânicas. Os computadores geram calor, mas não o suficiente para exigir refrigeração à base de água, o resfriamento por ar é suficiente. Eles produzem um zumbido baixo, quase inaudível durante a movimentação do dia, mas perceptível no silêncio da noite, a ponto de duas famílias vizinhas baterem à porta do data center para reclamar com alguma regularidade. O gerente, Javier Echeverria, parece constrangido ao admitir isso. Ele diz que está trabalhando com pesquisadores locais para reduzir o ruído e já modificou o sistema de refrigeração para torná-lo menos barulhento. Essa capacidade de resposta está muito distante da burocracia que o Mosacat enfrentou para chamar a atenção de uma empresa dos Estados Unidos.

A trinta minutos de carro, nos arredores da cidade, o governo delimitou uma grande extensão de terra para desenvolver um parque científico, o Parque de las Ciencias, que funciona como uma *zona franca*, uma área de livre comércio. Alguns o chamam de *zona América*,[85] com um toque de amargura, porque abriga principalmente empresas norte-americanas que não pagam impostos ao governo. O parque até lembra um pouco os Estados Unidos,[86] com os gramados exuberantes e bem-cuidados, o design simétrico e uma majestosa fonte em forma de relógio de sol que evoca a estética imponente do National Mall, na capital Washington. A página inicial do site do parque anuncia um país política e economicamente estável, com abundância de terra, energia e água. Foi assim que, em 2021, quando o GPT-3 despertou um novo interesse em modelos de IA de grande escala, o Google comprou 29 hectares de terra ali, 58 vezes o tamanho total do data center da Antel, para estabelecer uma nova sede para seu segundo data center na América Latina.

Contudo, o Uruguai não tinha água em abundância naquele momento. Assim como Chile, Iowa e Arizona, o país também enfrentava uma seca devastadora. A escassez era tão grave que agricultores estavam perdendo todas as suas colheitas, o que custou ao país mais de um bilhão de dólares em perdas

agrícolas; no verão de 2023, o governo de Montevidéu chegou a misturar água salgada contaminada no abastecimento de água potável da cidade.[87] As famílias que abriam as torneiras viam um líquido marrom e fétido, com um forte odor químico. Quem podia pagar comprava água engarrafada para beber e tomava banho com as janelas abertas para evitar inalar muitos carcinógenos. Quem não podia pagar bebia a água da torneira mesmo assim, o que levou muitas pessoas a sofrerem com dores de estômago, erupções cutâneas, agravamento de condições de saúde preexistentes e angústia de uma taxa crescente de abortos espontâneos.[88]

Após uma pandemia devastadora, muitas pessoas não tinham condições financeiras de comprar água engarrafada. Enquanto o Vale do Silício prosperava,[89] com as capitalizações do Google e da Microsoft ultrapassando 2 trilhões de dólares, impulsionadas em parte pelo aumento do trabalho remoto e da dependência de serviços em nuvem, os assentamentos ilegais no Uruguai cresciam exponencialmente. As *ollas*, cozinhas comunitárias, passaram a recusar crianças famintas. Os responsáveis por essas *ollas* também viviam na pobreza e mal conseguiam sobreviver. Fabiana, a animada chefe de uma *olla* que vive em um assentamento ilegal e é carinhosamente chamada de Reina Madre [Rainha Mãe],[90] fica em silêncio ao se lembrar desse período. "Ter que dizer: 'Não tenho nem um prato para o seu filho…'" Ela não termina a frase. "Foi horrível." Mesmo após uma vida inteira de pobreza, que incluiu varrer o chão de um bordel aos sete anos para sobreviver, ela considera a pandemia e os anos de seca como alguns dos piores momentos de sua vida.

A crise hídrica surgiu dos efeitos combinados das mudanças climáticas e da falha do Estado na alocação dos recursos hídricos: no Uruguai, mais de 80% da água doce são destinados à indústria, e não ao consumo humano, principalmente à agricultura comercial.[91] Isso inclui fazendas industriais de soja e arroz, além de plantações de árvores que abastecem a produção de papel. A maioria dessas fazendas não é administrada por empresas locais, mas por multinacionais que exportam o que produzem e demonstram pouca responsabilidade pelo meio ambiente uruguaio.[92] As operações exaurem os nutrientes do solo, dificultando ainda mais o cultivo de alimentos,[93] e poluem os cursos de água com um volume de fertilizantes tão elevado que o país se tornou um dos maiores consumidores *per capita* desse insumo no mundo, contribuindo para taxas anormalmente altas de câncer.

Daniel Pena, pesquisador de sociologia da Universidad de la República, em Montevidéu, que há anos estuda a política do extrativismo ambiental, estabelece uma conexão direta com a história colonial do Uruguai. Ele percorre o país em uma caminhonete velha para entrevistar agricultores e moradores dos bairros mais pobres, a fim de documentar de perto como eles são pressionados pela indústria.[94] Assim como no Chile, no Uruguai as multinacionais estrangeiras ainda ocupam um patamar acima das comunidades locais na hierarquia política. Durante a seca, a indústria continuou usando água sem restrições, retirando o que precisava diretamente do principal rio do

país, o Santa Lucía, que também abastece o sistema público de água de Montevidéu. Há duas décadas, após um forte movimento de ativismo ambiental, o Uruguai se tornou o primeiro país do mundo a reconhecer a água como um direito humano em sua Constituição. Hoje, em uma amarga ironia, foi a água potável, e não aquela para uso industrial, que sofreu os cortes mais severos durante a escassez.[95]

Então, quando o Google chegou, Pena ficou atento. Em verificações regulares no site do Ministério do Meio Ambiente do Uruguai, que lista os principais projetos industriais, ele se deparou com a proposta da empresa para o novo data center. Pena já havia lido sobre hiperescaladores que utilizavam água potável mesmo durante grandes secas e sobre o ativismo de comunidades como o Mosacat, que resistiam a esses projetos. No entanto, ao baixar os detalhes da proposta, encontrou os números relativos ao consumo de água marcados como confidenciais. Ele enviou um pedido de informação pública, algo que já havia feito com sucesso cerca de vinte vezes, mas o Ministério continuou a ocultar os dados, alegando sigilo. Pena se perguntou o que estavam escondendo e passou a se preocupar com o precedente que isso poderia estabelecer para outras empresas de nuvem, as quais inevitavelmente seguiriam o exemplo do Google e começariam a mirar o Uruguai para sua própria expansão. Foi então que ele recorreu à cláusula da água na Constituição. Com a ajuda de um amigo advogado disposto a trabalhar *pro bono*, Pena processou o Ministério.

Em março de 2023, quatro meses depois, Pena surpreendentemente venceu o caso. O ministro do Meio Ambiente revelou que o data center do Google planejava usar 7 milhões de litros de água por dia diretamente do abastecimento de água potável, o equivalente ao consumo diário de 55 mil pessoas.[96] Com grande parte de Montevidéu recebendo água salgada nas torneiras pouco tempo depois, as revelações foram explosivas. Milhares de uruguaios foram às ruas para protestar contra o Google e contra todas as outras indústrias que haviam levado o governo a desperdiçar os preciosos recursos de água doce do país. Os slogans de resistência ainda estavam rabiscados nas paredes da cidade e nas barreiras das estradas durante minha visita em junho de 2024. Isto não é seca, dizia um deles. É pilhagem.[97]

Pena vê os data centers da mesma forma que Ramos vê a mineração. O problema não é a infraestrutura em si, mas a escala em que o Vale do Silício está tentando construí-la. Foi essa escala que levou empresas como Google e Microsoft a se expandirem para o Chile e o Uruguai, mesmo enquanto esses países enfrentavam grave escassez de recursos. É essa escala que faz com que precisem de 58 vezes mais terra do que a Antel e operem com muito menos responsabilidade perante a população local. “São projetos extrativistas que vêm para o Sul para usar água barata, terras isentas de impostos e mão de obra muito mal remunerada. E então eles não contribuem para o nosso país; não melhoram nosso acesso à internet”, afirma.

No fim de 2023, o Google atualizou silenciosamente a proposta de

instalação do data center no Uruguai para adotar um sistema de resfriamento sem uso de água e reduzir o tamanho da instalação para um terço do previsto originalmente. Pena afirma que a luta está longe de terminar: hoje, o governo está ocultando o consumo de energia projetado da nova proposta, como segredo comercial. Ele também enviou ao Ministério uma petição, com mais de quatrocentos signatários, exigindo um estudo mais abrangente do impacto ambiental e social de toda a cadeia de abastecimento do data center: de onde estão sendo extraídos os minerais usados na produção do hardware, como a mão de obra envolvida é tratada, quanto carbono será emitido e de que modo o lixo eletrônico será descartado, para não contaminar a comunidade com produtos químicos. A maior parte desses impactos não ocorrerá no Uruguai, mas Pena sente solidariedade com os países onde isso inevitavelmente acontecerá. Eles são "geralmente todos do Sul", afirma. "Todos nós acabamos com as mesmas consequências, mas vindas de diferentes elos da cadeia de abastecimento global."

Em 2024, o tribunal ambiental do Chile decidiu que o Google não pode construir um data center que utilize água em Santiago. A porta-voz do Google Chile afirmou que a empresa continua comprometida com o país e com a América Latina, e que pretende iniciar o processo de licenciamento para um data center refrigerado a ar em Cerrillos, quando necessário.[98] Contudo, isso não impediu que outros hiperescaladores avançassem pela América Latina. Em 2022, a Microsoft definiu a localização de seu data center no Chile, logo após o segundo investimento na OpenAI — justamente na cidade natal de Arancibia, Quilicura.[99]

No que diz respeito aos desenvolvedores de data centers, a Microsoft foi uma iniciante tardia. Antes dos investimentos na OpenAI, o Google estava muito à frente em número de instalações ao redor do mundo. Contudo, com a explosão repentina da demanda por mais infraestrutura de computação para sustentar as ambições em IA, a Microsoft adotou o manual do Google e passou a segui-lo nas mesmas regiões.

O Chile aonde a Microsoft chegou era diferente do que havia recebido o Google. Em 2022, já tinham se passado mais de dois anos desde o *Estallido Social*, que deixou uma marca indelével na política do país. Após meses de protestos, o Chile passou por uma experiência notável: reescrever a Constituição, com ampla participação de cidadãos comuns, para substituir a que estava em vigor desde a ditadura brutal de Pinochet. No fim, as novas versões da Constituição não foram aprovadas; dois processos separados resultaram em dois documentos marcadamente partidários, que não conseguiram obter apoio amplo. No entanto, o processo reacendeu entre jovens e famílias da classe trabalhadora, em particular, um otimismo em relação à democracia. A ascensão das ideias e do apoio à esquerda levou, em uma reviravolta dramática, à eleição de um presidente *millennial* de esquerda: Gabriel Boric Font, de apenas 35 anos. No discurso de vitória, Boric repetiu

um dos slogans dos protestos, que criticava o legado dos Chicago Boys durante o governo Pinochet: "Se o Chile foi o berço do neoliberalismo, também será seu túmulo."[100]

Boric, que assumiu o cargo em março de 2022, também havia sido um estudante manifestante. Isso encorajou jovens de todo o país a usar o que aprenderam com a agitação social sobre organização e protesto para fundar uma nova geração de organizações ativistas progressistas. Elas passaram a fazer parte do ritmo cotidiano de cada comunidade, reunindo-se regularmente para imaginar com esperança o que desejavam para o futuro do Chile.

Foi nesse período que Arancibia fundou o próprio grupo de ativismo com outro jovem morador de Quilicura, Rodrigo Vallejos. Os dois se conheceram durante a organização da revolta social e rapidamente se uniram por uma profunda paixão pelo meio ambiente. Eles chamaram o grupo de *Resistencia Socioambiental Quilicura*, baseando-se no conceito amplamente estabelecido na América Latina de que o social e o ambiental são inseparáveis.

Após a chegada da Microsoft a Quilicura, Vallejos e Arancibia fizeram o que o Mosacat e Daniel Pena haviam feito: começaram a examinar minuciosamente todos os materiais públicos disponíveis sobre o projeto e a pesquisar tudo o que conseguiam encontrar. Vallejos, então estudante de Direito na Universidade Diego Portales, trabalhava até tarde da noite, entre as tarefas acadêmicas, lendo documentação técnica e aprendendo, por conta própria, como funcionam os data centers.

A projeção da Microsoft era que seria necessária uma quantidade significativamente menor de água do que o Google consumia em Cerrillos. Ainda assim, Vallejos estava preocupado. A seca no Chile só havia piorado, e as previsões indicavam que poderia durar até 2040. As zonas úmidas de Quilicura já sofriam enormemente, tanto pela invasão industrial quanto pela aceleração da desertificação. Em seu site, a Microsoft se gabava de inovações de ponta em sistemas de resfriamento para data centers, alegando que as instalações poderiam operar sem consumir água. Se a empresa tinha capacidade de construir data centers sem consumo hídrico, por que não estava fazendo isso em Quilicura?

"É impressionante que uma empresa com a reputação da Microsoft apresente publicamente um discurso ecológico, mas, na realidade, não cumpra os padrões globais de inovação num país emergente como o Chile", escreveu Vallejos em um artigo.[101]

A Microsoft explicaria, depois, que as inovações anunciadas ainda estavam em desenvolvimento e sendo testadas apenas em um único local nos Estados Unidos. "Então por que eles prometem essas coisas no site?", pergunta Vallejos.

Vallejos chamou a atenção de pesquisadores locais e internacionais, incluindo Marina Otero Verzier, diretora de pesquisa do Nieuwe Instituut, o instituto holandês de Arquitetura, Design e Cultura Digital, e Serena Dambrosio e Nicolás Díaz Bejarano, pesquisadores da FAIR, o think tank

coliderado por Martín Tironi Rodó.[102] Otero ficou comovida com a paixão e a luta incessante de Vallejos e Arancibia. Eles se desgastaram lendo os extensos documentos técnicos da Microsoft, escrevendo artigos críticos e protestando, mas enfrentavam grande dificuldade para conseguir uma audiência com a empresa ou com o governo. Ela começou então a refletir sobre como poderia ajudá-los. Como colocá-los na sala de negociação com as pessoas certas?

Marina Otero sabia que tinha algo que Vallejos não tinha: afiliações a universidades de prestígio como Harvard e Columbia, capazes de atrair a atenção da Microsoft e do governo. Ela montou uma campanha multifacetada, envolvendo-se tão profundamente que acabou deixando o próprio emprego. Passou a falar em conferências importantes sobre os impactos ambientais dos data centers e sobre a luta da *Resistencia* contra eles; estabeleceu conexões com o Ministério da Ciência, Tecnologia, Conhecimento e Inovação do Chile, além de representantes da Microsoft e do Google; e conectou Vallejos e Arancibia a outros pesquisadores internacionais para ampliar sua visibilidade.

Com Dambrosio e Díaz, Otero também desenvolveu um projeto mais especulativo. Os três tinham formação em arquitetura e vinham estudando a infraestrutura das tecnologias digitais modernas pelas lentes do ambiente construído. Eles começaram a se perguntar: e se tratassem os data centers como estruturas arquitetônicas e reimaginassem radicalmente a estética, o papel nas comunidades locais e a relação com o ambiente ao redor?

Díaz gostava de visitar bibliotecas nacionais durante suas viagens — lugares belíssimos, concebidos para expressar a grandeza das memórias e do conhecimento de um país. A ideia de que os data centers também pudessem ser considerados bibliotecas — com os próprios acervos de memória e conhecimento — chamou atenção. E, como ele argumentava, esses locais também poderiam ser projetados para serem acolhedores e belos, em vez de estruturas feias e extrativistas.

Isso representava um afastamento radical das definições de retribuição adotadas pela Microsoft e pelo Google, como os programas de impacto comunitário deste último, que Díaz descreve como iniciativas "esquizofrênicas", que tendem a ser desconectadas da forma como as comunidades são realmente afetadas pelas instalações das empresas. Com Vallejos e Arancibia, os três pesquisadores buscaram financiamento e organizaram um workshop de catorze dias, convidando estudantes de arquitetura de toda a Santiago para reimaginar como poderia ser um data center pensado para Quilicura.

Os alunos criaram maquetes impressionantes. Um grupo imaginou tornar o uso da água do data center mais visível, armazenando-a em grandes piscinas que também poderiam ser aproveitadas pelos moradores como espaço público. Outro propôs abandonar os projetos brutalistas típicos de data centers em favor de um "território de dados fluido", onde a infraestrutura digital coexistiria com os pântanos, mitigando os impactos nocivos. As estruturas do

data center também funcionariam como passarelas suspensas, convidando os residentes de Quilicura a percorrerem a área úmida e apreciar o ecossistema. Viveiros de plantas e estações de nidificação de animais seriam intercalados entre salas de servidores tradicionais, contribuindo para a reabilitação da biodiversidade local. O data center extrairia água poluída da zona úmida, purificaria para uso e a devolveria ao ambiente. Os próprios computadores coletariam e processariam dados sobre a saúde da região para acelerar a restauração do ecossistema. "Não estamos resolvendo o problema", afirma Díaz, "mas estamos imaginando outros tipos de relação entre os dados e a água."

No fim do workshop, os alunos apresentaram as ideias aos moradores e a outros integrantes da comunidade. "Foi uma conversa incrível", diz Otero. "É possível ver o quanto a comunidade tem conhecimento. Eles tinham muito a oferecer."[103]

Três anos após o início de seu mandato de quatro anos, Boric estava sob pressão para obter "vitórias rápidas" para a economia — ainda mais por ser um presidente jovem e de esquerda. Isso significava pressão para expandir a indústria de mineração e para viabilizar a chegada de 28 novos data centers. No dia em que Boric anunciou um plano para desenvolver uma estratégia nacional de data centers, Vallejos me enviou um vídeo da coletiva de imprensa acompanhado de um emoji: o rosto desorientado, com olhos em espiral.

Para ser justa, a coalizão formada por ativistas do norte do Chile e de Santiago, com pesquisadores nacionais e estrangeiros, deixou uma marca evidente.[104] Como parte do plano nacional para data centers, o Ministério da Ciência, responsável por supervisionar a implementação, formou, pela primeira vez, um comitê de ativistas para consultas regulares durante o processo de elaboração. Vallejos, Arancibia e membros do Mosacat foram convidados a integrar esse grupo. O Ministério também está coordenando um projeto de lei sobre IA que define como o Chile deseja abordar o desenvolvimento, a aplicação e a regulamentação da tecnologia. Antes, as discussões tratavam a IA como algo universalmente positivo; hoje, o tom mudou para reconhecer os custos sociais e ambientais envolvidos.

O Chile, como muitos países do Sul, aprendeu pela dura experiência que não deve esperar que o Norte decida como serão construídas suas tecnologias digitais. "A forma como construímos tecnologia responde a determinado quadro cultural e histórico", afirma Aisén Etcheverry, chefe do Ministério. Se a internet global foi moldada sem o Chile, o país tem a oportunidade de moldar a IA à sua maneira.

Martín Tironi Rodó vai um passo além. Para ele, é evidente que a indústria de IA hoje está enraizada em uma ideologia colonial: ela impõe ao planeta sua visão de mundo e sua tecnologia — o que é IA, o que é uma "boa" IA, o que significa criar uma indústria de IA. O Chile, afirma, poderia se tornar um líder na resistência a essa imposição. Após séculos de extrativismo,

o país compreende intimamente o que significa ter a terra saqueada, desapropriada e destruída em nome do progresso. Poderia usar essas experiências como fonte para gerar conceitos fundamentalmente novos — conceitos *decoloniais* — de inteligência artificial.

"No mercado planetário de IA, nós, como país, estamos desempenhando um papel específico: fornecer materiais para desenvolver a tecnologia", diz ele. "Muitas empresas estão tentando extrair inúmeros recursos de nós para criar IA."

"Portanto, precisamos pensar a partir dessa posição, dessa posição geopolítica, dessa posição terrestre. Podemos imaginar outra forma de interligar a inovação tecnológica com a Terra."

É uma ambição nobre, e as forças que se opõem a ela são poderosas.

CAPÍTULO 13

# OS DOIS PROFETAS

Em maio de 2023, Altman chegou a Washington, DC, para depor diante do Congresso. Foi uma performance extraordinária. Ele repetiu a promessa de que a IAG solucionaria as mudanças climáticas e curaria o câncer, apresentou um argumento convincente de que as tecnologias da OpenAI aprimorariam e criariam novos empregos "fantásticos", esquivou-se de perguntas sobre questões relacionadas a direitos autorais e sobre a falta de transparência e garantias de privacidade no que dizia respeito aos dados utilizados no treinamento, além de fazer um apelo sincero por um regulamento — quer dizer, um regulamento com a bênção da OpenAI, evocando o espectro da China para pedir aos legisladores que não desacelerassem sua inovação.

Os senadores o adoraram. Em uma interação reveladora que capturou a agilidade retórica de Altman e a confiança e o interesse que vinha conquistando, ele apresentou três recomendações de política que mudaram o foco da conversa sobre problemas existentes, como trabalho, meio ambiente e propriedade intelectual, para a regulamentação de sistemas futuros de IA e riscos extremos. A primeira foi sobre a criação de uma agência que desenvolvesse e administrasse um regime de licenciamento para modelos acima de determinado limite de capacidade; a segunda, sobre a criação de um conjunto de padrões de segurança de IA para medir capacidades "perigosas"; e a terceira, sobre a exigência de auditorias independentes desses padrões para verificar conformidade. Mais tarde, Altman detalhou que limites de capacidade poderiam ser aproximados por limites de computação, caso necessário, e que "capacidades perigosas" poderiam incluir a habilidade de um modelo manipular, persuadir e gerar receitas para novos agentes biológicos.

"Caso promulgássemos essas normas, você seria qualificado para administrá-las?",[1] perguntou o senador John Kennedy, da Louisiana.

"Eu adoro o meu trabalho atual", respondeu Altman, provocando risadas.

"Existem pessoas que seriam qualificadas para isso?", insistiu Kennedy.

"Ficaríamos felizes em recomendar alguns nomes, sim."

"Certo. Você ganha bastante dinheiro, não é?"

"Eu ganho… Não. Recebo o suficiente para o plano de saúde. Não tenho participação acionária na OpenAI."

Sentado ao lado de Gary Marcus, Altman conquistou até mesmo um de seus críticos mais ferrenhos. “Permita-me apenas deixar registrado que estou sentado ao lado de Sam, mais perto do que já estive dele, com exceção de uma única vez”, disse Marcus, “e a sinceridade dele… é muito evidente fisicamente, de um jeito que não é transmitido pela televisão.” (Mais tarde, Marcus retrocederia nessa rara demonstração de aprovação: “Percebi que eu, o Senado e, no fim das contas, o povo americano provavelmente fomos manipulados.”) [2]

A equipe de preparação de Altman considerou o episódio um sucesso absoluto.[3]

A audiência foi a cereja do bolo de uma longa campanha. Após o lançamento do ChatGPT, quase todos em Washington buscavam urgentemente reuniões com a OpenAI. A pequena equipe de políticas públicas liderada por Anna Makanju, depois de operar em relativa obscuridade, recebeu uma avalanche de pedidos. Por meses, com ou sem a presença de Altman, eles jantaram com autoridades, realizando demonstrações, respondendo perguntas e entregando a proposta legislativa apresentada durante o depoimento para o maior número possível de formuladores de leis — de senadores, deputados, assessores e membros do gabinete a diplomatas visitantes, chefes de agências e a vice-presidente Kamala Harris — do início da manhã até a noite, praticamente sem parar.[4]

De acordo com o *New York Times*, no início de junho, Altman já havia se reunido pessoalmente com ao menos cem legisladores dos Estados Unidos,[5] alguns dos quais chegaram a mencionar com orgulho essas conversas privadas durante a audiência.

No dia do depoimento de Altman, um pequeno grupo de artistas conceituais de Hollywood, especialistas na criação do design conceitual de personagens, objetos e outros elementos visuais para filmes, também tinha reuniões marcadas com diversos gabinetes do Congresso.[6] Eles haviam realizado um financiamento coletivo para arcar com as passagens e hospedagens e, durante esse processo, foram ameaçados por trolls na internet e tiveram seus dados vazados por se manifestarem contra a indústria de IA. Assim como os roteiristas de Hollywood — e, pouco depois, os atores —, estavam no meio de greves históricas, em parte para negociar por mais proteções contra a inteligência artificial. Os artistas também planejavam falar abertamente sobre os efeitos devastadores que a IA generativa já estava causando em sua profissão. Os desenvolvedores da IA generativa haviam treinado em milhões de obras de artistas sem o devido consentimento para criar produtos bilionários e empresas que, na prática, passaram a substituí-los. Esses empregos que estavam sendo eliminados eram empregos sólidos de classe média — talvez centenas de milhares.[7] “Os artistas estão sentindo uma dor excruciante no momento. Ninguém está conseguindo trabalho”, disse Karla Ortiz, artista conceitual conhecida por seu trabalho em *Doutor Estranho*, da Marvel Studios, que fazia parte do grupo e entrou com o primeiro processo

de artistas contra inúmeras empresas de IA generativa.

Quando chegaram a Washington, diversas de suas reuniões foram adiadas para o dia seguinte devido ao depoimento de Altman, bagunçando toda a agenda e deixando o grupo vagando pelos corredores da audiência — literalmente do lado de fora da sala onde tudo acontecia. No segundo dia, mais uma vez, o grupo precisou lutar por atenção. Altman estava participando de um jantar exclusivo com sessenta membros da Câmara no Capitólio, desfrutando de um bufê impecável que serviu frango assado.[8] Ao mesmo tempo, os artistas organizavam um happy hour interativo e tentavam atrair o maior número possível de assessores com o melhor que seu orçamento lhes permitia: vinho e fast-food.

Essa foi uma pequena ilustração, ainda sombria e quase cômica, de quem de fato detinha poder e influência na conversa sobre políticas de IA e de quem não os detinha.

A mesma narrativa que Altman[9] havia usado por muito tempo dentro da OpenAI para justificar a ocultação de pesquisas e a pressa nos avanços estava sendo habilmente empregada na época para direcionar o debate sobre a regulamentação da IA nos Estados Unidos para propostas que evitassem responsabilizar a empresa e, em alguns casos, consolidassem seu monopólio. A velha carta na manga do Vale do Silício que perguntava "Mas e a China?" continuava fazendo milagres para afastar a criação de normas. Naquele momento, ela estava sendo mais infalível do que nunca, com os receios de Washington em relação ao país asiático atingindo novos patamares, alimentado pela fenomenal ascensão do TikTok.

Esse medo podia ser exemplificado pelo clima no Departamento de Comércio, que havia se tornado a ponta de lança de uma ofensiva agressiva do governo norte-americano para sufocar o desenvolvimento da IA na China. No ano anterior, em 7 de outubro de 2022, o Departamento divulgou uma diretriz que afirmava ter como objetivo prejudicar os avanços militares chineses na inteligência artificial. Utilizando um mecanismo conhecido como "controles de exportação" — uma forma de delimitar a venda de certas tecnologias a países estrangeiros sob a justificativa que era por motivos de segurança nacional —, impôs, sem aviso prévio, restrições severas à exportação de chips de IA de última geração desenvolvidos nos Estados Unidos, principalmente os da Nvidia, para a China. As dimensões do impacto dessa medida foram muito além dos militares chineses; ela abalou as estruturas da comunidade científica e da indústria de IA chinesas, que trabalhavam em tudo, desde aplicativos para saúde e educação até equivalentes locais ao ChatGPT, que utilizava a tecnologia. "Se você tivesse me falado sobre essas normas cinco anos atrás, eu teria dito que se tratava de um ato de guerra — precisaríamos estar em guerra",[10] afirmou um analista de semicondutores.

Ao avaliar as consequências, o Departamento de Comércio parecia frustrado com a aparente resiliência da China. O ritmo de desenvolvimento e

a adoção da IA no país haviam desacelerado moderadamente, mas não o bastante para trazer tranquilidade. Parte disso se devia às próprias manobras da Nvidia:[11] a China representava um mercado gigantesco para a fabricante norte-americana de chips e, tão rapidamente quanto o Departamento havia imposto suas restrições, a empresa projetou novos chips que se encaixavam com perfeição nas regras para dar continuidade às vendas aos clientes chineses. A proibição também acabou impulsionando a indústria de fabricação de chips na China, que enfrentava dificuldades havia muito tempo para produzir componentes tão bons quanto os da Nvidia.[12] As ações do governo dos Estados Unidos geraram um aumento repentino no interesse por alternativas locais, garantindo ao setor um forte impulso de financiamento e de feedback para avançar.

Entretanto, o maior desafio aos esforços do governo vinha do vigoroso movimento transfronteiriço de inteligência artificial open-source, que estava rapidamente reproduzindo modelos corporativos fechados de IA generativa e os disponibilizando na internet para qualquer pessoa baixar e usar. Após correr atrás do prejuízo de maneira intensa,[13] a Meta se tornou uma força dominante ao oferecer gratuitamente seus modelos de linguagem de grande porte. Fazia muito tempo que a empresa defendia o desenvolvimento open-source; o cientista-chefe Yann LeCun acreditava na importância da ciência aberta. Além disso, também era uma jogada de negócios inteligente. A Meta não precisava vender modelos de IA generativa para lucrar, mas liberar modelos gratuitos, enquanto os integrava a seus principais produtos, poderia ajudá-la a consolidar sua liderança em inteligência artificial, atrair os maiores talentos científicos e provocar a concorrência por tais talentos que *de fato* dependiam da venda de modelos.

Tendo acumulado recursos computacionais similares aos que a OpenAI acumulara através da Microsoft, a Meta avançava a todo vapor na produção do Llama, seu equivalente à série GPT. Tecnicamente, o Llama não atendia à definição completa de open-source, que exigiria divulgar *tanto* os pesos do modelo *quanto* seus dados de treinamento. No entanto, o fato de ter cumprido apenas o primeiro requisito, publicando gratuitamente os pesos, foi o suficiente para transformar o Llama — apesar da política de não o disponibilizar na China — em um componente crucial para a indústria chinesa de IA.[14]

Em meio ao clima de frustração e medo em Washington, um livro branco de políticas que repetia as recomendações de Altman chegou dois meses depois de sua audiência, em julho de 2023.[15] Escrito por uma associação de pesquisadores, incluindo membros do clã de Segurança da OpenAI, da Microsoft e do Google DeepMind, além de mais de uma dúzia de think thanks, muitos conectados à comunidade dos pessimistas, o documento defendia novamente um novo regime de licenciamento para modelos de inteligência artificial baseado em limites de computação, e o desenvolvimento de avaliações de segurança em IA para capacidades perigosas, incluindo a

habilidade de manipular e convencer e a criação de novas receitas de armas biológicas. O documento de 51 páginas também deu um nome à categoria de modelos que precisariam dessa intervenção governamental: modelos de IA de fronteira. Segundo os autores, estes — modelos que poderiam apresentar tais capacidades perigosas — ainda não existiam, mas, ao ampliar modelos atuais de empresas como a OpenAI e a Anthropic com quantidades cada vez maiores de computação, eles poderiam surgir de repente e imprevisivelmente, a qualquer instante.

Depois da chegada do livro branco, em poucas semanas, a OpenAI e a Microsoft formaram uma aliança estratégica com o Google e a Anthropic para lançar o Frontier Model Forum, um grupo dedicado à promoção de pesquisas relevantes e à influência na agenda de políticas sobre riscos de segurança em IA. Era uma rara questão em que os interesses dos pessimistas, dos aceleracionistas e das empresas orientadas por lucro se alinhavam: manter os modelos de IA de fronteira no centro das discussões regulatórias do governo era ideologicamente primordial para os pessimistas e conveniente para os aceleracionistas e as corporações, já que desviava a atenção da regulamentação dos modelos de IA já existentes e de seus problemas. Todos também fortaleciam suas causas ao defender que os pesos dos modelos de ponta não fossem disponibilizados abertamente. No primeiro ano do fórum, a Meta estava perceptivelmente ausente. (Ela se juntaria no ano seguinte para garantir um assento à mesa.)

No centro das recomendações de Altman e da própria ideia de "modelo de IA de fronteira" estava a associação entre a escala de um modelo e as capacidades emergentes e, portanto, potencialmente perigosas. O argumento se baseava na filosofia das leis de escalabilidade — a ideia de que uma maior quantidade de computação de treinamento deveria, como esperado, gerar modelos mais potentes — e na crença, comum nos círculos pessimistas, de que uma IA extremamente avançada poderia sair do controle. As propostas de política derivadas desse pensamento giravam em torno do fato de reguladores basearem suas intervenções na quantidade de computação utilizada para treinar um modelo de aprendizagem profunda: aqueles que ultrapassassem determinado limite de computação deveriam automaticamente ser tratados com mais cautela e receber restrições mais severas.

O livro branco de políticas de julho de 2023 sugeriu um número para esse limite: $10^{26}$ operações de ponto flutuante, referindo-se ao número mínimo total de cálculos — um 1 seguido de 26 zeros — com que um modelo precisaria ser treinado para ser classificado como "de fronteira". Os autores admitiam que o limite era um tanto arbitrário, declarando apenas que era um nível de computação que os modelos existentes provavelmente ainda não haviam ultrapassado. Sara Hooker, vice-presidente de pesquisa da desenvolvedora de modelos de linguagem Cohere e uma das coautoras do documento, afirma que, conversando com os colaboradores que propuseram o

número, passou a acreditar que eles o escolheram para que fosse ligeiramente superior à quantidade de computação que a OpenAI teria utilizado para treinar o GPT-4.

Entretanto, Hooker e muitos outros pesquisadores, inclusive Deborah Raji, discordaram da abordagem de usar limites de computação para regular os modelos.[16] Embora a escala *possa* levar a capacidades mais avançadas,[17] o inverso não é verdadeiro: capacidades avançadas não exigem escalas. Por exemplo, um modelo de aprendizagem profunda treinado somente com dados biológicos de alta qualidade pode ser um gerador extremamente poderoso de receitas biológicas, mesmo em escala muito baixa. Por meio da destilação, uma das técnicas mencionadas pela OpenAI em seu plano de pesquisa de 2021, grandes modelos também podem ser transformados em modelos menores com capacidades similares, o que depende novamente dos dados de treinamento do modelo e do tipo de rede neural que está sendo treinada. No fim das contas, nem todos os modelos são baseados em Transformers. Sendo assim, de acordo com Hooker, a computação não é muito correlacionada com risco algum, muito menos com tipos específicos de riscos, como aqueles descritos no livro branco.

Sem consenso entre os coautores do documento sobre o sistema regulatório centrado em computação ou sobre o limite específico, o número de $10^{26}$ operações de ponto flutuante foi inserido em uma nota de rodapé e no apêndice, sem justificativa alguma, acompanhado de ressalvas importantes sobre o motivo de os limites serem uma abordagem extremamente imperfeita. O que chocou Hooker foi a rapidez com que não apenas o sistema, mas também o limite exato, se transformou em uma das propostas de política mais populares. A ideia conquistou uma enorme popularidade em Washington após o livro branco tocar diretamente nos receios em relação à China.[18] Modelos de IA de fronteira pareciam assustadores, ainda mais se Pequim tivesse acesso a eles. "Partes da administração estão tentando colocar as mãos em tudo que podem porque querem fazer *alguma coisa*",[19] afirmou Emily Weinstein, então pesquisadora do CSET, em uma conversa comigo no final de 2023.

As ideias do livro branco foram bem recebidas pelo público no Departamento de Comércio.[20] A equipe se mobilizou para se reunir com especialistas e discutir como o controle de modelos de IA de fronteira poderia ser feito e se seria viável mantê-los fora do alcance de Pequim. Em pouco tempo, passou a considerar uma proposta inédita: expandir seus controles de exportação de inteligência artificial para focar não somente hardware, mas também o próprio software, proibindo a exportação de modelos de IA acima de certo limite de computação. Em outras palavras, estava avaliando se poderia bloquear a publicação de pesos de modelos na internet e impedir que fossem amplamente oferecidos.

Perceptivelmente, uma recomendação crucial de Marcus e Christina Montgomery, vice-presidente da IBM, que também testemunharam ao lado de Altman, não chegou nem mesmo perto de ser tão popular, apesar de ambos

repetirem a ideia durante toda a audiência: obrigar as empresas a divulgarem exatamente o que compõe os dados de treinamento utilizados em seus modelos. Isso quase não ajudaria a impedir que capacidades avançadas chegassem a Pequim, segundo as preocupações de Washington, mas seria um passo decisivo para responsabilizar empresas por uma vasta gama de questões, incluindo o uso de materiais protegidos por direitos autorais, privacidade de dados de usuários e avaliações científicas rigorosas das capacidades dos modelos. "Se não sabemos o que há dentro deles, não sabemos ao certo o quão bem eles funcionam",[21] disse Marcus. Teríamos simplesmente que aceitar a palavra da empresa.

Essa abordagem também reduziria significativamente a incerteza sobre se e como capacidades perigosas poderiam surgir, pela mesma razão que faz com que a computação seja um mau previsor de risco: o comportamento de um modelo de aprendizagem profunda deriva, antes de tudo, de seus dados. Se um desenvolvedor cria um modelo de linguagem capaz de gerar receitas de armas biológicas, "é porque foi treinado com um conjunto de dados que incluía informações sobre armas biológicas",[22] afirma Sarah Myers West, codiretora-executiva do AI Now Institute e ex-conselheira sênior sobre IA na FTC. Como sempre, a rede neural está apenas revelando padrões presentes nos dados de treinamento; sendo assim, tornar esses dados acessíveis seria o primeiro passo para estabelecer clareza científica sobre quais tipos de dados podem levar a resultados perigosos.

À medida que o Departamento de Comércio consultava diversos especialistas sobre sua proposta de restrição da divulgação de pesos de modelos,[23] a notícia dessas deliberações, que seria anunciada no início de 2024 mediante uma audiência pública, dividiu a comunidade de desenvolvimento e a de políticas de IA, que crescia rapidamente, em dois blocos. O embate era Fechado contra Aberto, tecnonacionalismo contra ciência sem fronteiras. Além dos participantes do Frontier Model Forum e da comunidade pessimista em geral, o lado cujo voto era Fechado logo conquistou o sistema de segurança nacional e inteligência dos Estados Unidos. Do outro lado estavam a Meta, desenvolvedores de IA open-source, startups, grupos da sociedade civil e acadêmicos independentes.[24]

Enquanto o lado Fechado continuava enfatizando muitos dos mesmos pontos que executivos da OpenAI já vinham sinalizando internamente por anos, o lado Aberto argumentava que isolar modelos seria muito mais prejudicial do que benéfico. O empecilho para produzir novas armas biológicas, por exemplo, não está em encontrar uma receita, observou Weinstein. Elas já existem em abundância[25] na internet e são facilmente encontradas no Google. A inconveniência está em obter os materiais e equipamentos necessários para de fato fabricar esses armamentos. A restrição de acesso a supostos modelos de IA de fronteira, portanto, seria pouco útil para resolver essa situação. No entanto, os danos colaterais de suprimir a

publicação de modelos de IA poderiam enfraquecer as bases da inovação em inteligência artificial nos Estados Unidos. O open-source, ou seja, o compartilhamento e o desenvolvimento com base em códigos e softwares disponibilizados para a comunidade, sempre foi o alicerce sobre o qual prosperaram as inúmeras startups norte-americanas. Bloquear a publicação de pesos de modelos daria aos desenvolvedores menos caminhos do que nunca para criarem os próprios produtos e serviços de IA. Isso fortaleceria ainda mais o domínio das gigantes representadas no Frontier Model Forum.

O exame detalhado de modelos de inteligência artificial também se tornaria cada vez mais difícil, como no trabalho de Sasha Luccioni, Yacine Jernite e Emma Strubell, que dependem fortemente de modelos generativos abertos para quantificar os custos de carbono e os impactos ambientais da escala contínua desses sistemas.

Essencialmente, ao contrário[26] de representar um risco à segurança nacional, um enorme nível de colaboração aberta entre fronteiras foi justamente o que fortaleceu a liderança norte-americana em IA. Como os dois países que mais produzem talento e pesquisa em inteligência artificial no mundo, os Estados Unidos e a China há muito tempo são vistos como os principais colaboradores mútuos no desenvolvimento da tecnologia. Por mais de uma década, cientistas e empreendedores de ambos os países têm desenvolvido sobre o trabalho uns dos outros, acelerando o avanço do campo e de uma ampla gama de aplicações muito mais rapidamente do que cada grupo conseguiria sozinho, beneficiando não apenas as duas nações, mas inúmeros outros países ao redor do globo. Um dos exemplos mais famosos[27] é o da ResNet, uma das redes neurais mais utilizadas no mundo, que foi publicada por pesquisadores chineses no escritório da Microsoft em Pequim. Ela não somente sustenta grandes sistemas de visão computacional e de reconhecimento de voz e linguagem, como também foi um ingrediente central da primeira versão do AlphaFold, da DeepMind — um sistema de IA lançado em 2018 capaz de prever a estrutura tridimensional de uma proteína a partir de sua sequência de aminoácidos, algo primordial para acelerar o desenvolvimento de medicamentos e a compreensão de doenças. (Os avanços posteriores da DeepMind no AlphaFold, utilizando outra rede neural, renderiam a Dennis Hassabis e a outro pesquisador sênior do laboratório o Prêmio Nobel de Química de 2024.)

Ainda assim, em outubro de 2023, as ideias defendidas pelo lado Fechado receberiam sua maior validação até o momento ao aparecerem na ordem executiva sobre IA do governo de Biden.[28] A ordem, uma das mais longas da história, parecia um amontoado de documentos escritos por grupos totalmente diferentes — porque foi exatamente isso que aconteceu.[29] Um deles tinha suas raízes no *Blueprint for an AI Bill of Rights*, de 2022, que a Casa Branca havia elaborado com cuidado ao longo do tempo, em consulta com grupos da sociedade civil, para delinear como a IA poderia ser desenvolvida, utilizada e controlada de formas que fortalecessem os direitos civis, a justiça racial e a

proteção de privacidade. Entre outras coisas, enfatizava desenvolver a inteligência artificial com ampla participação de comunidades e especialistas, enfrentar o impacto discriminatório da IA em áreas como saúde e contratação de pessoal e proteger as pessoas da coleta não autorizada de dados.

O outro documento era uma reprodução surpreendentemente fiel das recomendações de Altman e da estrutura apresentada no livro branco sobre modelos de IA de fronteira, anexado de última hora após o texto chamar a atenção de algumas pessoas que simpatizavam com a ideologia pessimista dentro da Casa Branca. O livro branco havia enfatizado a necessidade de focar modelos futuros de IA, que ainda não existiam; a ordem executiva dedicaria, mais adiante, metade de seu conteúdo a tais modelos. Além disso, o livro branco descrevia quatro exemplos de capacidades perigosas; a ordem executiva manteria três delas: a geração de novas receitas de armas QBRN (químicas, biológicas, radiológicas e nucleares), ataques cibernéticos automatizados e, em uma réplica literal, a evasão do controle humano "por meio de fraude e ofuscação".

Para o espanto de Hooker, Raji e muitos outros pesquisadores de IA, o limite exato do livro branco, de $10^{26}$ operações de ponto flutuante, também apareceu na ordem executiva como o ponto acima do qual modelos precisariam ser reportados ao governo dos Estados Unidos.

Com um aval desse porte, a abordagem baseada em limites de computação rapidamente se propagou. No fim do ano, ela seria adotada na Europa, que escolheu o limite de $10^{25}$ para algo ligeiramente mais restritivo, enquanto legisladores que pressionavam pela aprovação do tão aguardado EU AI Act se sentiam atropelados pelos repentinos avanços da IA generativa e procuravam às pressas formas de incorporá-los à lei. No início de 2024, a abordagem chegaria à Califórnia com a introdução do SB 1047, um novo projeto de lei de segurança em IA, que retornaria ao limite de $10^{26}$. O governador Gavin Newsom acabaria vetando o projeto,[30] o que, ironicamente, a OpenAI e outros desenvolvedores de modelos haviam pressionado bastante para que acontecesse, devido ao enorme conjunto de outras propostas de responsabilização presentes no texto. Muitos criticaram Newsom por ceder aos interesses da indústria; no entanto, para diversos pesquisadores, incluindo Hooker e Raji, o veto foi um alívio. "Foi um passo na direção certa para garantir que estamos ancorados no consenso científico", declarou Hooker. "Ainda temos perguntas importantes a serem respondidas sobre o motivo deste número e quais riscos exatamente se pretende evitar."[31]

Toda a sequência de eventos — o depoimento de Altman, o livro branco, a intensa campanha de influência política, a hiper-reatividade de Washington aos receios em relação à China e a rápida consagração dos limites de computação em documentos de política pública nos Estados Unidos e no exterior — foi uma demonstração contundente, entre outras coisas, de quanto a expertise independente em inteligência artificial havia atrofiado. No mês anterior, em setembro de 2023, Raji se viu como a única acadêmica sem

vínculos financeiros com a indústria ou a comunidade pessimista a testemunhar diante do Congresso ao lado de Altman, Musk, Nadella, Gates, Zuckerberg, Pichai e Jack Clark, dentre outros executivos da área tecnológica.[32] Todos estavam presentes no primeiro dos AI Insight Forums do senador Chuck Schumer, uma das séries de reuniões sobre políticas públicas mais disputadas e decisivas daquele ano para dar início à criação de uma legislação sobre IA. Enquanto seus colegas de painel faziam afirmações espetaculares e infundadas sobre as promessas e os perigos da inteligência artificial, pontuadas por referências convenientes à necessidade de "superar a China", que deixavam os senadores ainda mais em alerta, o que mais surpreendeu Raji foi ver como muitos na plateia pareciam acreditar em tudo aquilo.[33]

Ela percebeu que as pessoas sentadas ao seu lado, e suas enormes equipes de políticas públicas, monopolizaram a narrativa em Washington por tanto tempo que muitos formuladores de políticas públicas, a partir de então, a tratavam como verdade absoluta. Um porta-voz de Schumer diria mais tarde à imprensa que o senador estava consultando pessoalmente Altman e outros executivos da OpenAI à medida que avançava rumo à regulamentação.[34] "Para mim, isso foi uma grande revelação", disse Raji. "Caramba, precisamos de mais gente apenas desmentindo — apenas olhando para o que as pessoas estão afirmando, e dizendo: 'Na verdade, a realidade é mais complicada.'"

Washington foi apenas o auge da etapa norte-americana da charmosa ofensiva política de Altman. Em março daquele ano, depois de publicar um tweet sobre seus planos para viajar ao exterior e se encontrar com usuários, a viagem evoluiu para uma odisseia por diversas cidades e continentes, em que ele posou para fotos com praticamente todos os presidentes do G20.[35] Passou até mesmo a ter uma nova marca: a Turnê Mundial de Sam Altman.

Não havia nenhuma grande estratégia por parte das equipes de comunicação ou de políticas públicas da OpenAI por trás da turnê. Altman simplesmente escolheu suas primeiras paradas e as anunciou para seu mais de 1,5 milhão de seguidores no Twitter. Após as primeiras visitas, o interesse saiu rapidamente do controle, e as equipes de comunicação e políticas foram envolvidas na confusão. A cada novo trecho da viagem, apressavam-se para organizar a logística, já se preparando para as dificuldades que, com certeza, surgiriam devido à falta de planejamento.

Era a manifestação de uma dinâmica que sempre havia estado presente: Altman seguindo o próprio caminho. Por vezes, esse caminho se alinhava com o da empresa; por outras, não. Durante o lançamento do GPT-4, a OpenAI havia elaborado cuidadosamente todos os anúncios e materiais de divulgação para apresentar o projeto como o esforço coletivo que realmente havia sido. Mais de cem funcionários estavam envolvidos no desenvolvimento do modelo.[36] O autor do anúncio oficial da empresa era simplesmente "OpenAI".[37] Altman então fez um tweet dando crédito a uma única pessoa: Jakub Pachocki.[38] O cientista realmente havia desempenhado um papel

importante, mas era um entre dezoito líderes do projeto. Alguns funcionários se perguntaram se sua contribuição tinha sido realmente singular. Altman reforçou a narrativa ao escolher Pachocki para se sentar atrás dele durante sua audiência no Senado, em Washington.

Essa dinâmica também se mostrava de outras formas: havia executivos oficiais na empresa, mas nem sempre ficava claro se eram eles que Altman realmente ouvia, nem se eram os processos e decisões formais ou seus relacionamentos e impulsos pessoais que guiavam a estratégia da companhia.

À medida que a OpenAI se especializava depressa e recebia mais exposição e escrutínio, essa incoerência nas camadas mais altas se tornava cada vez mais significativa. A empresa já não era apenas as divisões de Produtos e de Pesquisa. Naquele momento, existiam diversos departamentos voltados ao público: além da equipe de comunicação, havia uma equipe jurídica que redigia pareceres e lidava com um número crescente de processos. A equipe de políticas públicas estava se espalhando por vários continentes. Cada vez mais, a OpenAI precisava comunicar uma narrativa única e coerente a seu público, além de definir posições claras para poder articulá-las. No entanto, em diversas ocasiões, a falta de clareza estratégica resultou em mensagens públicas confusas.

No fim de 2023, o *New York Times* processou a OpenAI e a Microsoft por violação de direitos autorais, acusando-as de treinar seus modelos com milhões de artigos do jornal. No início de janeiro, a resposta da OpenAI, redigida pela equipe jurídica, tinha um tom excepcionalmente combativo e acusava o *Times* de "intencionalmente manipular nossos modelos" para gerar evidências que sustentassem tal argumento.[39] Na mesma semana, a equipe de políticas públicas da empresa[40] enviou uma declaração ao Comitê de Comunicações e Digital da Câmara dos Lordes, no Reino Unido, afirmando que seria "impossível" para a OpenAI treinar seus modelos de ponta sem materiais protegidos por direitos autorais. Após a imprensa destacar a palavra *impossível*, a empresa logo deixou de usá-la.

"Simplesmente há muita confusão, o tempo todo", afirmou um funcionário de um departamento voltado ao relacionamento com o público. Embora parte disso reflita os típicos obstáculos de crescimento de startups, o perfil e o alcance da OpenAI já ultrapassavam em muito a fase relativamente inicial em que a empresa ainda se encontrava, acrescentou ele. "Não sei se existe uma prioridade estratégica na alta liderança. Sinceramente, acho que as pessoas só tomam as próprias decisões. E então, de repente, elas ganham um tom estratégico, mas, na realidade, não passam de acidentais. Às vezes, não há um plano; apenas o caos."

Enquanto Altman viajava mundo afora, pegando voos para a Europa, a América Latina, o Oriente Médio, a Ásia e a África, encantando — e em apenas algumas ocasiões ofendendo — audiências cuidadosamente selecionadas de estudantes, investidores de tecnologias e fãs, a falta de clareza

estratégica intensificava as antigas fraturas da OpenAI e acelerava a empresa rumo a extremos mais opostos do que nunca.

De um lado, a divisão de Produtos continuava no controle, correndo contra um número jamais visto de concorrentes para instituir as tecnologias da OpenAI mais rápido do que nunca. Naquele momentos, contava com o reforço das outras divisões em expansão e de muitas pessoas da divisão de Pesquisa, energizadas pela crença — alimentada pelo aumento dramático do alvoroço e das expectativas — de que desenvolver e lançar os modelos da empresa era a melhor forma de cumprir sua missão.

Do outro lado, o clã de Segurança, espalhado pela divisão de Pesquisa e ainda com muitos membros nas equipes de pesquisa de política de Miles Brundage e de alinhamento de Jan Leike, passou a ser uma minoria evidente. Eles compensavam essa considerável desvantagem numérica alertando mais veementemente do que nunca sobre as capacidades perigosas e os riscos existenciais que acreditavam que poderiam surgir a qualquer instante. Conforme os modelos da OpenAI continuavam a avançar, alguns pesquisadores que antes não se identificavam com o clã de Segurança passaram a abraçar suas ideias, à medida que o rápido aumento das capacidades os convencia de que a IA poderia atingir um nível de inteligência capaz de corromper o controle humano e passar a ter um comportamento errático. Desse lado do espectro, o mesmo aumento drástico do entusiasmo e das expectativas significava que a empresa tinha uma obrigação moral de agir com máxima cautela para cumprir sua missão.

Eram os pessimistas e os aceleracionistas encarnados — dentro da própria OpenAI.

A separação chegava até a liderança. Depois do DALL-E e do ChatGPT, a maioria dos executivos e gerentes seniores havia se sentido cada vez mais confortável com a ideia de os modelos serem ferramentas benéficas disponibilizadas no mundo por meio de uma "implementação iterativa", expressão criada pela OpenAI entre lançamentos para descrever sua nova abordagem. Diferentemente do lançamento escalonado do GPT-2 ou da liberação controlada via API do GPT-3, a implementação iterativa significava ir com tudo — entregar os modelos aos usuários o quanto antes e com frequência. E, como em todas as estratégias de implementação anteriores, a empresa apresentou um novo argumento explicando por que *essa* era a abordagem mais segura possível. Segundo Altman e outros executivos, a implementação iterativa[41] garantiria às pessoas e instituições um período de adaptação, ao mesmo tempo que permitiria testar os modelos nas mãos de usuários reais, coletar feedback real e aprimorar os produtos. "Acho que não terminaria bem desenvolver um sistema de IA superpoderoso em segredo e, então, lançá-lo ao mundo de uma vez só", afirmou Altman em seu depoimento no Senado.

Contudo, se havia um líder na OpenAI que não estava totalmente de acordo com isso, era Sutskever. À medida que os modelos avançavam e o

impacto de seus lançamentos se acelerava, ele acreditava que a empresa precisava aumentar, ao invés de diminuir, a cautela quanto ao potencial desses sistemas de culminar em consequências devastadoras. Depois do GPT-4, Sutskever, que antes dedicava a maior parte do tempo a desenvolver as capacidades dos modelos, fez uma virada brusca e se voltou para a segurança de IA. Começou a dividir seu tempo pela metade. Para aqueles ao seu redor, às vezes parecia estar em pé de guerra consigo mesmo. Era simultaneamente aceleracionista e pessimista: mais empolgado e mais assustado do que nunca com a chegada da IAG e a possibilidade de ela superar rapidamente os humanos e se tornar uma superinteligência.

Sutskever passou a falar em tons cada vez mais messiânicos, intrigando até amigos de longa data e preocupando outros funcionários. Durante uma reunião com um novo grupo de pesquisadores, ele expôs seus planos para se preparar para a IAG.

"Quando todos nós entrarmos no bunker…", começou ele.

"Perdão", interrompeu um pesquisador. "No bunker?"

"Certamente iremos construir um bunker antes de lançarmos a IAG", respondeu Sutskever, como se fosse algo óbvio. Estava evidente que uma tecnologia tão perigosa se tornaria objeto de intensa cobiça de governos do mundo inteiro. Poderia amplificar tensões geopolíticas; os principais cientistas que estivessem trabalhando na IAG precisariam ser protegidos. "Claro", acrescentou, "vai ser opcional entrar ou não no bunker."

O pesquisador continuou tendo grande respeito por Sutskever, mas também manteve certa distância. "Há um grupo de pessoas — Ilya é uma delas — que acredita que criar a IAG trará uma espécie de arrebatamento. Literalmente, um arrebatamento", afirmou.

À medida que Sutskever continuava se dividindo quanto ao posicionamento, uma nova ideia começou a se difundir entre ele e Altman: uma equipe focada inteiramente no desenvolvimento de novos métodos de alinhamento para superinteligências, antecipando quais métodos, como a aprendizagem por reforço a partir de feedback humano, deixariam de ser suficientes quando os sistemas pudessem, em sua visão, superar os humanos. Altman a chamou de Projeto Manhattan de Alinhamento. A princípio, eles conversaram sobre a criação de uma organização separada: uma entidade independente, sem fins lucrativos, com um fundo inicial de um bilhão de dólares. Como Sutskever não queria deixar a OpenAI, e devido a questões de acesso aos modelos, ambos decidiram manter o projeto dentro da empresa. A OpenAI então anunciou em um blog a formação de uma nova equipe para conduzir a ideia, assim como seu novo nome: Superalignment. A publicação também comunicava um compromisso ambicioso: dedicar 20% da capacidade computacional que a empresa havia garantido até o momento à nova equipe. [42] Sutskever e Leike seriam os líderes do projeto.

Em outra reunião, Sutskever se colocou diante dos funcionários para apresentar a nova equipe e seus objetivos. Pegou o microfone e começou a dar

batidinhas no aparelho: *tum, tum, tum.*

"O alinhamento é uma chama ardente", disse. "O Superalignment é um incêndio devastador."

Não muito tempo depois, com a empresa já grande demais para o Mayo, o escritório decorado com plantas e fontes, para onde haviam se mudado após a pandemia, os executivos transferiram a maior parte da divisão de Pesquisa de volta para o antigo Pioneer Building. A mudança basicamente dividiu as duas metades da empresa — Produtos e Segurança — em mundos distintos.

Em julho de 2023, pouco após a OpenAI tornar pública a criação da equipe do Superalignment, a empresa alugou um cinema no Metreon, no centro de São Francisco, para que os funcionários assistissem ao filme *Oppenheimer*, que conta a história do físico J. Robert Oppenheimer como líder do Projeto Manhattan, que visava à criação da primeira arma nuclear do mundo.

"eu estava esperando que o filme do Oppenheimer inspirasse uma geração de crianças a se tornarem físicas, mas deixou a desejar nesse aspecto",[43] publicou Altman no Twitter. "vamos fazer esse filme acontecer! (acho que a *A rede social* conseguiu inspirar os fundadores de startups.)"

Por quase oito anos, a analogia que Altman fizera em seus primeiros e-mails para Musk, comparando a OpenAI ao Projeto Manhattan, permaneceu como um escopo persistente dentro da empresa, sendo usada até mesmo nas orientações de novos contratados. Ele gostava dela.[44] Fazia aniversário no mesmo dia que Oppenheimer, informação que costumava salientar a repórteres. Também gostava de parafrasear a crença do criador da bomba de que "a tecnologia acontece porque é possível".[45] Além disso, sempre parecia omitir que o físico passou a segunda metade da vida mortificado pelo arrependimento e fazendo campanha contra a disseminação da própria criação.

Funcionários diferentes atribuíam significados distintos à analogia. A maioria via o Projeto Manhattan como um acontecimento histórico, que representava a capacidade de realizar um avanço tecnológico capaz de salvar o mundo e de mudar a história antes que inimigos perigosos tivessem a oportunidade, concentrando talento e recursos em uma escala exorbitante. Entre o clã de Segurança, a analogia evidenciava a gravidade do fardo que a OpenAI carregava ao apresentar uma tecnologia que poderia comprometer a própria existência da humanidade.

Para Altman, ela representava uma lição de relações públicas. "A maneira como o mundo foi apresentado à energia nuclear é uma imagem que ninguém jamais vai esquecer, aquela nuvem em formato de cogumelo sobre o Japão",[46] dissera anos antes em um evento. "Pensei muito sobre o motivo pelo qual o mundo se voltou contra a ciência, e uma das respostas dentre as inúmeras em que estou disposto a acreditar é aquela imagem, além da ideia de que talvez alguma tecnologia seja poderosa demais para estar ao alcance das pessoas. Elas se convencem mais por imagens do que por fatos."

Aqueles no extremo espectro da segurança em IA, com os níveis mais altos de probabilidade de desastre, ficavam cada vez mais inquietos com o otimismo aparentemente simplório com que Altman e o restante da empresa encaravam essa história. Em certo momento, a OpenAI também organizara uma exibição de *Apollo 11*, um documentário sobre o programa Apollo, para levar o primeiro homem à Lua, outra analogia de que o Vale do Silício gostava. "Por que falar sobre o Projeto Manhattan se podemos simplesmente dizer 'Progama Apollo'? Por que trazer todo esse peso?", contestou um pessimista radical.

Uma cena em particular de *Oppenheimer* ficou gravada em sua mente: o momento que antecedeu o teste Trinity, a primeira detonação de uma bomba atômica, quando o físico, interpretado pelo ator Cillian Murphy, calcula que as chances de o mundo ser destruído são "próximas de zero".

"*Próximas* de zero?", responde o major general Leslie R. Groves, vivido por Matt Damon, incrédulo.

"O que você quer apenas da teoria?", retruca Oppenheimer.

"Zero seria bom", diz Graves.

"Essa é uma analogia à situação em que nos encontramos", disse o pessimista radical. "Ainda não podemos de maneira alguma calcular qualquer coisa sobre essas IAs."

À medida que a OpenAI continuava avançando em sua pesquisa, para alguns, a analogia com o Projeto Manhattan começou a adquirir um novo significado.

Com a desaceleração no avanço dos modelos de linguagem de grande porte, devido ao esgotamento de dados e da computação, a divisão de Pesquisa passou a apostar mais fortemente no desenvolvimento de agentes de IA. A ideia, que ganhava força em todo o campo, era uma retomada do debate entre as hipóteses de "linguagem pura" e "fundamento". A linguagem pura estava atingindo o limite, assim como a combinação entre linguagem e visão. Para muitos na comunidade de inteligência artificial, a impressão era de que a próxima etapa de avanço provavelmente dependeria de agentes capazes de executar ações no mundo real e coletar feedback do ambiente. Dentro da OpenAI, essa capacidade também era vista como uma maneira de obter vantagem competitiva. Um assistente de IA que pudesse conversar com o usuário era bom, mas um que fosse capaz de automatizar tarefas complexas, como enviar e-mails ou programar sites, era ainda melhor. Isso não somente tinha uma enorme relevância comercial, mas também poderia acelerar o progresso interno da própria empresa.

O esforço mais ambicioso dentro da divisão de Pesquisa era o AI Scientist, uma tentativa de construir um agente que fosse capaz de realizar pesquisas científicas de modo independente. Com poucas informações "novas", ou seja, poucos dados restantes para serem extraídos da internet ou de livros, os pesquisadores do projeto tinham grandes esperanças de que um "cientista" autônomo pudesse gerar o próprio conhecimento executando experimentos. A equipe havia sido formada a partir da fusão entre o antigo time de geração de

código, que trabalhava no Codex, e outro time que tentava resolver o desafio do raciocínio utilizando grandes repositórios de problemas matemáticos e suas soluções como um conjunto de dados estruturado para ensinar lógica passo a passo ao modelo. Ambas as capacidades — programar e resolver problemas matemáticos — pareciam blocos de construção promissores para realizar experimentos e analisar dados.

Na liderança do projeto estavam Jakub Pachocki e Szymon Sidor, que buscavam criar não apenas um cientista, mas, especificamente como Altman desejava, um pesquisador de IA autônomo, que ajudaria a OpenAI a exceder seus avanços em inteligência artificial.

O projeto fez com que aqueles que acreditavam na IAG se sentissem ao mesmo tempo extremamente empolgados e nervosos: se o AI Scientist fosse bem-sucedido, a IAG certamente seria alcançada mais rápido. Isso encurtava a linha do tempo tanto para a utopia quanto para a possível eliminação da humanidade. Assim como Sutskever, alguns pesquisadores começaram a mencionar "um bunker" em conversas casuais, imaginando até algo semelhante a Los Alamos: em alguma área remota do deserto dos Estados Unidos, uma equipe de elite de pesquisadores de IA viveria e trabalharia em instalações seguras para protegê-los de ameaças externas.

Em 2023, eles acreditavam que tais ameaças incluíam ataques direcionados e até uma IAG descontrolada. No Slack, a equipe de segurança postou um rascunho de seu modelo de ameaças para a OpenAI, muito mais desenvolvido do que na época em que os executivos ainda debatiam sobre quão rígida deveria ser a segurança da organização. No documento, constavam três categorias de ameaças: atores estatais estrangeiros, concorrentes e pessoas motivadas por ideologias.

Na terceira categoria, o rascunho exibia um link para um artigo de Eliezer Yudkowsky, um pessimista radical e líder na comunidade de segurança em IA, que criou e popularizou o termo *friendly AI* para se referir a sistemas bem alinhados, além de ter escrito uma obra de fanfic muito estimada chamada *Harry Potter and the Methods of Rationality*. O livro, com 122 capítulos e mais de 660 mil palavras, reimagina Harry interagindo com o mundo bruxo como um racionalista bem treinado. Para muitos, o texto serviu como porta de entrada para o altruísmo eficaz e, mais tarde, para a ideologia pessimista mais ampla. Yudkowsky também havia sido um dos fundadores do blog *LessWrong*, um hub central para pesquisadores de segurança em IA se conectarem e divulgarem ideias da área, onde ele vinha defendendo, com pânico crescente, uma pausa total no desenvolvimento da inteligência artificial, à medida que a probabilidade de desastre subia para 95%. Em março de 2023, escreveu um artigo para a revista *Time* no qual falava sobre a dor de ver a filha perder o primeiro dente e se perguntar se ela teria chance de crescer.[47] Além disso, propôs um plano para forçar a interrupção dos avanços em IA: o fechamento de todos os grandes *clusters* de GPU, o rastreamento das vendas de GPUs e, se necessário, um ataque com bombas a "um data center fora de controle". O

rascunho do modelo de ameaças da OpenAI continha o link para esse artigo como exemplo de um ideólogo que defendia a violência.

Um pequeno grupo de funcionários da OpenAI que admirava as ideias menos violentas de Yudkowsky considerou extremamente frustrante o fato de o rascunho citá-lo, mas não mencionar a própria IA hostil como ameaça. Após receber o feedback, a equipe de Segurança adicionou uma quarta categoria de ameaça: sistemas de IAG desalinhados.

À medida que a OpenAI era catapultada a novas alturas, o conselho vinha perdendo integrantes sem repô-los. Desde o início de 2023, três diretores independentes haviam deixado a empresa em rápida sucessão, em parte devido à corrida frenética que o ChatGPT desencadeou para construir e comercializar as tecnologias de IA generativa em toda a indústria.

Em fevereiro, Reid Hoffman foi o primeiro a deixar o conselho, após cinco anos, devido a conflitos de interesse. No ano anterior, havia fundado uma startup, a Inflection, com Mustafa Suleyman, hoje ex-cofundador da DeepMind, que estava rapidamente se tornando uma concorrente direta da OpenAI.

Um mês depois, o exemplo de Hoffman foi seguido por Shivon Zilis, a substituta de confiança de Musk e diretora da Neuralink, que, após o desligamento de Musk, continuou supervisionando a OpenAI em nome dele e entrou oficialmente no conselho em 2020.[48] Fazia muito tempo que alguns líderes temiam que Zilis transmitisse informações sensíveis da empresa a Musk, mas ela havia se comprometido a manter a confidencialidade em relação à OpenAI acima de sua lealdade ao ex-copresidente rejeitado. Em julho de 2022, tal posição tornou-se extremamente questionável quando veio à tona que ela tivera gêmeos com Musk, sem revelar a informação aos demais diretores.[49] Ainda assim, Altman tentou mantê-la no conselho por motivos que os outros membros não compreendiam, chegando a procurar a orientação de Zilis, em outubro de 2022, sobre como lidar com a aparente irritação de Musk com a escalada da avaliação da OpenAI, que já estaria se aproximando de 20 bilhões de dólares. "Isso é uma fraude",[50] Musk escrevera para Altman, após mencionar suas contribuições substanciais ao financiamento inicial da empresa. Em março de 2023, finalmente o papel de Zilis no conselho tornou-se insustentável quando Musk incorporou a xAI, uma nova empresa de inteligência artificial e mais uma concorrente direta da OpenAI.

O terceiro a se afastar foi Will Hurd, ex-representante republicano do Texas e ex-agente da CIA, que havia ingressado no conselho em 2021. Quando sua nomeação foi anunciada, Altman disse aos funcionários que era importante ter alguém que equilibrasse o viés liberal do Vale do Silício. Ele então organizou uma reunião para que qualquer pessoa que tivesse ressalvas pudesse fazer perguntas a Hurd. Os funcionários não se reprimiram, questionando-o sobre suas opiniões a respeito de diversos temas, incluindo Donald Trump. Em junho de 2023, Hurd deixou a OpenAI para se

concentrar integralmente em sua campanha presidencial nos Estados Unidos. Em outubro, ele desistiria da corrida.

Ao lado de Altman, Brockman e Sutskever, restavam apenas três membros independentes no conselho: Adam D'Angelo, cofundador e CEO do Quora; a roboticista Tasha McCauley; e Helen Toner, pesquisadora do CSET.

Entre os três, D'Angelo havia sido o primeiro a entrar. Ex-colega de ensino médio de Zuckerberg no internato Phillips Exeter, D'Angelo tinha exercido por dois anos o cargo de CTO do Facebook antes de fundar o Quora. Em 2014, o Quora entrou para a Y Combinator na primeira leva de grupos sob a presidência de Altman; em 2017, um ano antes da nomeação de D'Angelo para o conselho, Altman aumentou o investimento da YC no Quora, coliderando uma rodada de financiamento de 85 milhões de dólares. No anúncio, Altman elogiou D'Angelo como um dos "CEOs mais inteligentes do Vale do Silício".[51] "E ele tem um foco de longo prazo, algo que se tornou um produto raro atualmente nas empresas de tecnologia", declarou Altman.

McCauley entrou para o conselho mais tarde, em 2018. Empreendedora que cofundara uma startup de robótica de telepresença e dirigia uma empresa de simulação urbana em 3D, ela havia se conectado com Altman por meio de Alan Kay, seu mentor, e conhecia Holden Karnofsky. McCauley era bastante respeitada na comunidade de segurança em IA e integraria o conselho do Centre for the Governance of AI, uma organização de pesquisa na área e, por um tempo, o conselho da Effective Ventures Foundation, uma companhia sediada no Reino Unido que supervisiona o *80,000 Hours*, um podcast popular do movimento AE. Karnofsky, que fazia parte do conselho da OpenAI na época, a indicou em seu esforço para encontrar mais diretores independentes enquanto a empresa se preparava para realizar a transição para uma estrutura de lucro limitado.

Toner foi a última adição ao conselho, em meados de 2021, também por indicação de Karnofsky. Dessa vez, suas recomendações eram de candidatos para substituí-lo, já que seu mandato de três anos estava chegando ao fim, junto com a formação da Anthropic. Antes de se estabelecer como especialista em tecnologias emergentes e na China, ela havia trabalhado com Karnofsky na GiveWell e na Open Philanthropy. Chegou aos temas de segurança em IA por intermédio do movimento AE, mas aos poucos foi se afastando dele. Em seu post mais popular no EA Forum antes do colapso da FTX, Toner observou que o movimento estava se tornando cada vez mais dogmático e socialmente isolado, e comentou que estava "tendendo ao desencanto com o AE".[52] Ainda assim, continuava sendo muito respeitada nesses círculos e dedicava seu tempo à comunidade mais ampla de segurança em IA, integrando com McCauley o conselho do Centre for the Governance of AI.

No final do verão de 2023,[53] fazia meses que o conselho se via em um impasse sobre quem nomear como novos diretores independentes. Como

parte do esforço para aumentar a supervisão após a demonstração do GPT-4, e ainda mais depois do lançamento do ChatGPT, McCauley havia coordenado um processo de cerca de um ano, incluindo entrevistas com funcionários e partes interessadas fora da empresa, para definir como um conselho reformulado e mecanismos de supervisão mais profissionalizados deveriam ser.

Durante o processo, o conselho, incluindo Altman, havia concordado que, devido à crescente importância dada às capacidades da empresa, o próximo diretor independente precisaria ter uma formação profunda em segurança de IA. Sendo assim, passaram meses compilando uma lista de candidatos e entrevistando cinco deles, incluindo Dan Hendrycks, uma figura central na comunidade pessimista, diretor do Center for AI Safety, em Berkeley, e o único conselheiro da xAI de Musk. No entanto, quando os membros independentes do conselho tentaram avançar para escolher um dos cinco, Brockman e Sutskever fizeram diversas objeções em relação aos candidatos avaliados. Como sempre, Altman mantinha uma postura discreta: não discordava abertamente de nenhuma opção, mas também não procedia com nenhuma delas. Em inúmeras ocasiões, sugeriu novos candidatos, todos com a mesma característica: pertenciam à sua rede e estavam, financeiramente ou de outras formas, dentro de sua esfera de influência.

Quando se tratou de estabelecer os novos mecanismos de supervisão, que incluíam diferentes canais para aumentar a visibilidade do conselho sobre as práticas de segurança e proteção da empresa, os diretores independentes também tiveram a sensação de que isso não era uma prioridade de Altman. No início do mandato de McCauley como diretora, ele a havia designado intermediária e representante dos funcionários no conselho; sendo assim, ela passou a se reunir regularmente com eles, enquanto parte do seu horário de trabalho. Em certa ocasião, também levou seu marido, o ator Joseph Gordon-Levitt, a um encontro externo da empresa, no qual ele assistiu atentamente a apresentações técnicas. Durante a pandemia, porém, essas reuniões foram minguando. Mais tarde, McCauley continuou mantendo alguns encontros regulares, mas a contagem das horas trabalhadas nunca voltou a acontecer.

Sem uma forma sistemática de se conectar com os funcionários, as informações sobre os acontecimentos na empresa acabavam chegando aos diretores independentes por meio de seus próprios relacionamentos pessoais nas comunidades de segurança em IA e tecnologia. Eles também dependiam do próprio Altman como canal para acompanhar informações importantes. O que começou a preocupá-los cada vez mais foi o quanto a retórica de Altman muitas vezes divergia de outros relatos que estavam ouvindo. Enquanto ele frequentemente apresentava um quadro otimista, os diretores recebiam cada vez mais, pelas próprias fontes, informações sobre inúmeros problemas, incluindo a falta de preparação da empresa antes do ChatGPT e a turbulência considerável após, as contínuas preocupações de segurança ligadas ao lançamento do GPT-4 e o ritmo extraordinário com que a OpenAI se

apressava para lançar novos produtos antes de solucionar muitos de seus problemas.

Um episódio pareceu particularmente óbvio. No fim de 2022, o conselho havia realizado um encontro presencial — o primeiro do que deveria ser uma reunião anual — durante o qual Altman salientou os fortes protocolos de segurança e testes que a OpenAI tinha executado com o Comitê de Segurança de Implementação (CSI) para avaliar a efetivação do GPT-4. Depois da reunião, um dos diretores independentes conversava com um funcionário quando ele mencionou que uma violação dos protocolos do CSI já havia sido cometida.[54] A Microsoft fizera um lançamento limitado do GPT-4 para usuários na Índia, sem a aprovação do CSI. Apesar de ter passado um dia inteiro reunido com o conselho naquele encontro, Altman não mencionou a violação uma única vez.

Embora os diretores independentes do conselho não tivessem motivo para acreditar que algo perigoso havia sido lançado ao público, eles ficaram inquietos com o aparente descaso com que a Microsoft havia quebrado o protocolo — e com o fato de Altman ter ignorado o incidente. Os modelos da OpenAI estavam em uma trajetória de avanço acelerado e, do ponto de vista da segurança de IA, eles acreditavam que logo poderia chegar um momento em que uma violação como aquela resultaria em consequências potencialmente catastróficas, se não existenciais. Na visão deles, a negligência de Altman em relação à falha da Microsoft estabelecia um precedente arriscado para a maneira como ele trataria os processos de segurança dos modelos quando os riscos aumentassem.

Enquanto isso, havia outros exemplos preocupantes do comportamento de Altman. Em março de 2023, ele enviou um e-mail ao conselho, sem copiar D'Angelo, declarando que acreditava que havia chegado a hora de D'Angelo renunciar. Argumentou que o fato de o Quora estar desenvolvendo o Poe, seu próprio chatbot, representava um conflito de interesses. A alegação pareceu repentina e de motivação duvidosa. Toner, McCauley e D'Angelo haviam, em diferentes momentos, feito perguntas inoportunas a Altman sobre as práticas de segurança da OpenAI, a solidez da entidade sem fins lucrativos e outros assuntos. Com a diminuição de seus aliados no conselho, Altman parecia, aos olhos dos três, estar procurando um pretexto para afastar um deles. Em resposta ao e-mail de Altman, um dos diretores independentes contestou. O Poe era um conflito de interesse moderado se comparado ao que Altman permitira por anos a fio no caso de Hoffman e Zilis. A tentativa de Altman falhou, e D'Angelo permaneceu no cargo.

Pouco tempo depois, D'Angelo estava em um jantar quando ouviu que o Startup Fund da OpenAI tinha uma estrutura estranha. O fundo oferecia aos investidores acesso antecipado aos produtos da OpenAI, um tipo de tratamento preferencial que deveria ser reservado aos próprios investidores da empresa. Depois de ouvir isso pela segunda vez, os diretores independentes pressionaram Altman para terem acesso a documentos sobre a estrutura do

fundo. Quando ele finalmente os entregou, os diretores descobriram que a estrutura não era apenas estranha; Altman era o *proprietário* legal do fundo, quando este deveria pertencer à OpenAI.[55]

Para os diretores independentes, cada episódio constituía um quadro único e preocupante: pouco a pouco, Altman tentava ofuscar a visibilidade do conselho e agir de maneira a não ser contido. Durante anos, havia promovido a capacidade do conselho de equilibrar seu poder, e até mesmo de demiti-lo, como o mecanismo de liderança mais importante da OpenAI. Na época, ele ostentava essa capacidade mundo afora para conquistar a confiança do público e dos governos.

No outono, o moral dos diretores independentes chegou ao fundo do poço, enquanto eles lutavam para fazer qualquer progresso significativo na negociação para a entrada de um novo membro do conselho. A cada mês que passava, essas lacunas na liderança se tornavam mais urgentes. A OpenAI estava prestes a treinar o GPT-5 e fazia avanços no AI Scientist. Então, de repente, no início de outubro, Toner recebeu um e-mail inesperado: Ilya Sutskever queria conversar.

CAPÍTULO 14

# LIVRAMENTO

Nas semanas que antecederam a abordagem de Sutskever a Toner, Altman estava lidando com uma crise de relações públicas. Depois de anos fora da mídia, a história sobre o afastamento de sua irmã e o fato de ela ter começado a trabalhar no ramo sexual finalmente veio à tona.

Em 25 de setembro de 2023, Elizabeth Weil, jornalista que escrevia perfis na revista *New York*, publicou uma matéria sobre Sam Altman que, pela primeira vez na grande imprensa, mencionava a existência de Annie. Comparou detalhes da vida dela, incluindo seus recorrentes problemas de saúde e o fato de viver com grande dificuldade financeira, sem a segurança de uma moradia, com o estilo de vida de Sam, dono de casas multimilionárias e carros de luxo. "Annie Altman?", escreveu Weil em seu artigo.[1] "Os leitores do blog de Altman, de seus tweets, de seu manifesto *Startup Playbook*, assim como das centenas de artigos sobre ele, saberão muito bem quem são Jack e Max… no entanto, Annie não existe na vida pública de Sam. Ela jamais faria parte do clube."

Altman sabia que os detalhes estavam prestes a serem revelados. No período que precedeu a publicação, que ocorreu durante o Yom Kippur, o feriado judaico do perdão, Weil havia dado à OpenAI a oportunidade de comentar sua matéria; a *New York* também tinha trabalhado com a empresa no processo de verificação de fatos. A tarefa de facilitar essa resposta e tentar controlar a história recaiu sobre Hannah Wong, vice-presidente de comunicações da OpenAI que, de repente, se viu atuando como intermediária entre a revista e a família de Altman, perguntando-se se aquilo realmente deveria fazer parte de seu trabalho.

No último dia antes da publicação da matéria, quando já estava claro que os detalhes sobre Annie seriam incluídos, ela recebeu um e-mail de Sam.[2] "oi annie. no espírito de estarmos quase no yom kippur, eu gostaria de pedir desculpas e seu perdão por uma coisa", escreveu. Durante os pedidos de ajuda da irmã, Altman havia se sentido no meio do fogo cruzado, dividido entre querer seguir a orientação da mãe, concordar com o restante da família sobre a necessidade de Annie aprender a ser independente financeiramente e sentir

que ela precisava de ajuda médica e passava por dificuldades para sobreviver. "ainda assim, tomei a decisão errada e deveria simplesmente ter simplesmente continuado [*sic*] a te ajudar; peço desculpas sinceras", disse.

A guinada brusca em relação à recepção pública extremamente positiva que Altman vinha recebendo desde o lançamento do ChatGPT, além de o fato de ver aquilo que considerava questões familiares dolorosas e privadas sendo expostas, pesaram sobre ele. A repercussão da matéria piorou tudo. Após a publicação, tweets antigos de Annie, alegando que Sam havia abusado dela de diversas maneiras, ressurgiram e começaram a viralizar. No entanto, por mais pesado e desafiador que fosse para Altman, para Annie aquilo representou uma extraordinária catarse. Durante anos, além do estresse crescente causado por seus problemas de saúde e pela falta de estabilidade, ela sentia como se estivesse gritando para o nada e tendo sua existência apagada.

Em 2024, eu a procurei para compreender melhor seu lado da história.[3] Também entrei em contato diretamente com a mãe, Connie Gibstine, os irmãos Max e Jack, e com Sam por meio da equipe de comunicações da OpenAI, a fim de ouvir suas versões. Annie colaborou avidamente, cheia de esperança de, por fim, contar com uma plataforma para compartilhar experiências pessoais que, segundo ela, eram fundamentais para o entendimento do caráter de Sam. Forneceu uma série de correspondências com a família, registros de saúde física e mental desde a infância até a vida adulta e outras evidências que validavam tanto o rompimento com a família quanto a deterioração de suas condições de vida.

Gibstine ofereceu uma breve declaração enfatizando o amor da família pela filha e a preocupação com seu bem-estar, ao mesmo tempo que negava as acusações de Annie, que chamou de "horríveis, profundamente dolorosas e falsas".[4] Gibstine recusou-se a ter uma conversa mais aprofundada ou a responder detalhadamente perguntas que buscavam sua perspectiva sobre o relato da filha. Não respondeu a meus pedidos extras de documentos que comprovassem as próprias afirmações. Max e Jack não responderam, e a OpenAI não comentou detalhes específicos.

Em janeiro de 2025, depois de Annie entrar com uma ação judicial contra Sam, ele, Gibstine, Max e Jack divulgaram uma nota de repúdio mais contundente sobre as acusações, caracterizando-a como uma pessoa mentalmente instável e excessivamente exigente quando o assunto era dinheiro, e afirmando que suas acusações "evoluíram de forma drástica ao longo do tempo".[5] "É dilacerante demais quando ela se recusa a receber tratamento convencional e ataca os membros da família que estão genuinamente tentando ajudar", disseram.

A partir de minhas conversas com Annie e da documentação concedida por ela, surgiu um quadro complexo da jornada conturbada que a levou a tornar públicas suas acusações no Twitter e, mais adiante, no processo judicial, assim como os detalhes de sua vida compartilhados para a matéria de Weil. Não tive acesso ao raciocínio completo por trás de muitas das decisões de sua

família, e a veracidade de algumas de suas declarações, em particular o suposto abuso sexual cometido por Sam na infância, não é comprovada, mas sua história se tornou, para mim, um microcosmo dos diversos temas que definem a história mais ampla da OpenAI. Também me ajudou a solidificar o entendimento do quanto a empresa é um reflexo e uma extensão do homem que a comanda. Os esforços obstinados de Annie para acrescentar sua perspectiva ao registro rapidamente se transformaram em um problema para a OpenAI. A cobertura sobre a irmã afetava Sam talvez mais do que qualquer outro tipo de matéria, o que acabou arrastando Hannah Wong, na função de chefe de comunicação da empresa, para as tentativas de conter a disseminação da história. Isso também se tornou um problema de outra maneira: na época em que as acusações de Annie vieram à tona pela primeira vez, outros executivos da OpenAI estavam prestando atenção.

De diversas formas, Annie e Sam, com nove anos de diferença, são surpreendentemente parecidos. Muitas das palavras usadas por pessoas próximas a ele para descrevê-lo também se aplicam claramente à irmã: uma excelente ouvinte, alguém que se lembra dos menores detalhes sobre os outros, muito brincalhona, extremamente generosa e rápida em ganhar a confiança das pessoas.

Como o irmão, Annie também se destacou nos estudos. Além de Sam, foi a única a se formar na Burroughs, impressionando James Roble, seu professor de física, que ainda se recorda dela com carinho mais de uma década depois, como uma aluna talentosa e de temperamento radiante.[6] Na época de faculdade, cursou a Tufts University, onde se formou em biopsicologia e fez uma habilitação secundária em dança. Após a graduação, em 2016, completou os pré-requisitos de medicina, antecipando a possibilidade de um dia ingressar no curso, e, em seguida, mudou-se para a Bay Area para trabalhar em um laboratório de neurociência na Universidade da Califórnia, em São Francisco. Em outras palavras, antes de sua vida sair dos trilhos, ela era o perfil típico de uma jovem ambiciosa, com ótima formação e inúmeras possibilidades pela frente.

Entretanto, as circunstâncias da vida de Annie começaram a se deteriorar após uma série de desafios inesperados. Em 2014, ainda na faculdade, recebeu o diagnóstico de tendinite de Aquiles, uma inflamação no tendão causada por esforço excessivo, e um esporão ósseo, o que a forçou a usar uma bota ortopédica.[7] Esse foi o primeiro item de uma lista crescente de problemas de saúde que passariam a atormentar seu corpo com dores crônicas e, em seus piores momentos, limitar severamente sua mobilidade e afetar sua qualidade de vida. Em um período de seis anos, Annie enfrentou amidalites, dores pélvicas e crises de tendinite, todas recorrentes, que a fizeram repetir o uso de botas ortopédicas, se espalharam além do tornozelo direito e, por vezes, dificultavam até mesmo ficar de pé por curtos períodos, além de um número crescente de cistos ovarianos que culminaram em um diagnóstico de síndrome

de ovários policísticos, ou SOP, que vez ou outra a fazia acordar com os lençóis encharcados de suor, um sintoma comum.[8]

Em meio a todos esses problemas de saúde, Annie sofreu um golpe ainda mais paralisante: no dia 25 de maio de 2018, seu pai, com quem tinha a relação mais próxima na família, faleceu de um infarto fulminante.[9]

Annie enfrentou dificuldades em relação à saúde mental ao longo de toda a vida. Ainda jovem, recebeu os diagnósticos de transtornos de ansiedade generalizado e obsessivo-compulsivo, precisando tomar Zoloft por quase uma década até que uma psiquiatra a ajudou a interromper o uso durante a faculdade.[10] Antes do falecimento do pai, ela estava muito melhor. Havia adotado ferramentas não psiquiátricas para a saúde mental, como meditação, e buscava formas de tratar a inflamação principal que parecia estar por trás da reincidência de seus problemas físicos, abraçando por completo uma alimentação saudável, orgânica e baseada em alimentos integrais. Sua jornada de saúde despertou um novo interesse em medicina alternativa, e ela se formou como professora de ioga. Também estava se apoiando em seu amor de longa data pela arte e havia escrito o rascunho de um livro que intitulou de *The Humanual*, reunindo suas reflexões sobre a vida e sobre como ser uma pessoa boa. Todavia, a morte do pai a destroçou, fazendo sua saúde mental sair do controle, e parece ter marcado uma aceleração na deterioração da saúde física. No final de 2018, já fazia parte de sua rotina consultar-se com inúmeros médicos e especialistas.

Sam também se referiu publicamente ao falecimento do pai como o pior momento de sua vida,[11] que se deu poucos meses depois de Musk deixar a presidência da OpenAI e Sam assumir integralmente. Pessoas próximas a ele dizem que a perda o abalou e, por um tempo, ele se comportou de modo errático devido à dificuldade que teve para lidar com a situação. A morte do pai foi visivelmente um ponto de inflexão, tanto na relação entre Annie e Sam quanto na relação dela com o restante da família. Segundo Annie, sua conexão com eles, em especial com a mãe, já estava tensa por conta de divergências sobre inúmeras decisões de vida, incluindo seu afastamento de uma carreira médica tradicional e a escolha de interromper o tratamento com Zoloft para seguir com meios naturais de tratamento. O pai era o único a lhe dar apoio incondicional; sem ele, Annie se viu isolada da família, mas ao mesmo tempo ansiava por aceitação e ajuda. Em vez disso, após confrontos por causa de dinheiro, a relação chegou ao ponto de ruptura.

Após o falecimento do pai, em 2018, Annie estava morando em Los Angeles. Trabalhou meio período como redatora assistente e, depois, em um dispensário de maconha. Ela diz que herdou cerca de 100 mil dólares do seguro de vida do pai. Empenhou-se ainda mais na tentativa de transformar seu amor por arte e performance em uma carreira séria: fez aulas de comédia, participou de noites de microfone aberto, criou um podcast e reescreveu o primeiro rascunho de seu livro como um espetáculo solo, que rebatizou de *The*

*HumAnnie*. Em meados de 2019, os fundos estavam se esgotando, a maior parte, segundo ela, gasta nos diversos investimentos artísticos, no aluguel caro, no plano de saúde pago do próprio bolso e em despesas médicas que se acumulavam — exames de imagem, fisioterapia, psicoterapia e corridas de Lyft e Uber para comparecer às consultas.

Para Annie, foi o que aconteceu em seguida que marcou o início do fim de sua relação com a família. Em maio de 2019, enquanto a dor crônica persistia, ela soube que o pai havia deixado para ela seus fundos de aposentadoria.[12] Sendo assim, Annie deixou o trabalho no dispensário e elaborou um plano de seis meses para usar aquela margem financeira extra — um pouco mais de 40 mil dólares — para cuidar da própria saúde e colocar em prática seus projetos criativos, o bastante para, com sorte, gerar uma fonte de renda sustentável. No entanto, pouco depois, sua mãe, que tinha controle dos fundos de aposentadoria do pai por ser a cônjuge sobrevivente, notificou a filha de que, na verdade, ela não receberia o dinheiro. Gibstine escreveu à Annie em um e-mail que a melhor estratégia tributária seria manter os fundos em seu próprio nome, em uma conta com imposto diferido que seria transmitida a ela por meio de um trust, ao qual teria acesso aos 59 anos e meio.[13] "Todos nós desejamos o melhor para você, e acreditamos que o melhor para você (e para todos) é a independência financeira, que traz satisfação em longo prazo, crescimento pessoal e segurança", disse Gibstine. "Por esse motivo, queremos esclarecer que não iremos ajudá-la financeiramente se/quando a herança atual do seu pai acabar." O e-mail foi assinado por "Sua mãe e seus irmãos Sam, Max e Jack".

Annie estava fragilizada. As anotações de sua terapeuta a respeito do mesmo período dizem que a medida tomada pela família, com a esperança de fazê-la se reerguer, acabou piorando sua condição.[14] Em dezembro de 2019, a conta bancária de Annie ficou no negativo.[15] Anteriormente, no mesmo ano, Sam havia conquistado o primeiro investimento de um bilhão de dólares da Microsoft para a OpenAI, e a empresa avançava a todo vapor no treinamento do GPT-3. Assustada e sozinha, Annie se cadastrou em um serviço de acompanhantes chamado SeekingArrangement e mostrou os seios para um homem em uma videochamada em troca de dinheiro suficiente para fazer sua conta sair do vermelho.[16]

Annie diz que nunca havia pedido ajuda financeira à mãe ou aos irmãos. Ela não presumia que poderia contar com a riqueza deles, mas também não acreditava que seria deixada sem uma rede de apoio caso realmente se visse em uma situação de emergência. Entre o fim de 2019 e meados de 2020, Annie fez diversos pedidos à família, necessitando de ajuda financeira, inclusive após a pandemia se tornar uma dose extra de estresse e incerteza.[17] Depois que Sam e a mãe participaram de duas sessões de terapia familiar com ela, eles concordaram em cobrir suas despesas por parte do ano.[18]

Em suas conversas com a família, fica evidente que eles se preocupavam que o dinheiro fosse um facilitador para comportamentos prejudiciais e

acreditavam que a melhor maneira de ajudar Annie a reestabelecer sua saúde mental seria incentivando-a a ser financeiramente independente. Tanto a declaração de Gibstine para mim quanto a versão pública que a família apresentou afirmaram que suas ações foram guiadas por aconselhamento profissional sobre o que seria melhor para Annie.

Em maio de 2020, quando o período de ajuda de custos da família se aproximava do fim e Annie continuava enfrentando dificuldades, ela pediu mais auxílio para pagar a fisioterapia e a fonoaudióloga, que, segundo contou a eles, custavam respectivamente 45 e 15 dólares por sessão.[19] Eles negaram e disseram que acreditavam que ela deveria arcar sozinha com as despesas de junho, inclusive com o depósito de aluguel que ela receberia de volta. Annie juntou seus poucos pertences e viajou para o Havaí, onde estivera quando o pai fizera sua última visita poucos meses antes de morrer. Encontrou um trabalho em troca de moradia em uma fazenda, com tarefas físicas leves, como capinar e plantar. Sam lhe enviou um e-mail pedindo o novo endereço. Oito meses depois do falecimento do pai, ele tinha solicitado a cada irmão que enviasse à mãe uma mecha de cabelo para ser misturada às cinzas do pai e transformada em um diamante para cada um deles.[20] "Será majoritariamente carbono vindo do corpo do nosso pai, mas terá um pouquinho de cada um de nós também", escrevera. O diamante de Annie estava pronto, e Sam queria enviá-lo.

Para Annie, o e-mail pareceu um tapa na cara. Ela não queria um diamante caro; queria que ele a ajudasse a garantir comida e moradia. Não sabia dizer se Sam simplesmente não entendia ou não se importava com a gravidade de sua situação. O abismo entre ela, ele e o restante da família rica parecia irreconciliável. Com imensa dor e luto, Annie cortou contato com todos.

Ao longo dos três anos seguintes, antes do contato de Elizabeth Weil, a vida de Annie atingiu o fundo do poço. Ela enfrentou insegurança em relação à moradia, alimentação e cuidados de saúde e recorreu ao trabalho sexual, virtual e presencialmente, para conseguir pagar as contas. Em sua declaração pública, a família disse que tentou "de diversas maneiras apoiar Annie e ajudá-la a encontrar estabilidade". Na versão dela, uma dessas iniciativas de apoio ocorreu na primavera e no verão de 2021, quando Sam tentou se reconectar com a irmã. Falaram-se por telefone três vezes, recorda-se Annie, durante as quais ele expressou o quanto a amava e se disponibilizou para lhe comprar uma casa. Annie e a família descrevem tal oferta de maneiras divergentes. Annie diz que não era para ela ter a propriedade em seu nome, mas sim para residir nela, um arranjo que a jovem entendeu como uma maneira de a impedir de vender o imóvel e que, em sua visão, poderia ser mais uma forma de Sam e a mãe imporem suas vontades sobre as próprias decisões dela em relação à saúde e à carreira. De acordo com Gibstine, em sua declaração para mim, a família ofereceu a propriedade da casa para Annie, mas a mãe também

não respondeu a meus pedidos por documentos que provassem isso. Na versão pública da família, eles afirmaram que se prontificaram a comprar uma casa para ela por meio de um trust, para que Annie pudesse ter um local onde morar sem a possibilidade de vendê-lo imediatamente. No fim, Sam e a irmã chegaram a um acordo distinto: ele pagaria seu aluguel por um ano diretamente ao proprietário. No verão de 2022, quando ela não retomou contato, os pagamentos foram interrompidos.

A história de Annie aprofunda os retratos conflitantes que as pessoas fazem de Sam. Ele é, ao mesmo tempo, generoso e interesseiro, acolhedor e ameaçador, um benfeitor para tantas e a fonte de uma dor pessoal profunda para outras. Alguém que projeta sinceridade e altruísmo em público, mas revela um cálculo mais complicado em seus comportamentos a portas trancadas. Alguém que pode dar e tirar, deixando muitos com a impressão de que fazem parte de um grande jogo de xadrez cujo tabuleiro completo somente ele consegue ver, e cujo objetivo final é preservar seu poder como rei.

A história de Annie também complica a grandiosa narrativa que Sam e outros executivos da OpenAI têm construído sobre a inteligência artificial, de que ela inaugura um mundo de abundância. Altman já disse que espera que a IA dê um fim à pobreza. Brockman repete, em suas histórias sobre o amigo e sobre Anna, sua esposa, que a IAG aprimorará drasticamente o sistema de saúde. Sutskever afirma que a tecnologia trará uma psicoterapia incrivelmente eficaz e baratíssima. No entanto, diante da realidade da vida dos trabalhadores no Quênia, dos ativistas no Chile e da própria experiência da irmã de Altman carregando o peso de todos esses problemas, tais sonhos soam vazios.

Apesar de todos os avanços impressionantes nas capacidades de IA, nenhum deles ajudou a aliviar o desespero de Annie. Se algo mudou, foi para pior: a inteligência artificial talvez tenha contribuído para aprisioná-la ainda mais. Ela não queria recorrer ao trabalho sexual. Segundo a jovem, aquele, era o último de seus recursos. Quando a dor crônica se intensificou e dificultou demais até mesmo o trabalho em troca de moradia, Annie tentou primeiro meios digitais para monetizar sua arte. Continuou com seu podcast e manteve uma loja no Etsy, além de uma conta no Patreon, mas nenhuma das opções rendia o suficiente nem para custear sua conta de celular.[21] Algo estranho estava se passando, e ela documentou isso em capturas de tela ao longo do tempo: recebia pouca ou nenhuma visibilidade em todas as suas redes sociais.[22] Às vezes, percebia que blocos inteiros de avaliações de seu podcast no app da Apple desapareciam misteriosamente, diminuindo suas chances de ser descoberta.[23] Ao menos duas vezes, tanto no Instagram quanto no YouTube,[24] acumulou visualizações e, em seguida, as perdeu misteriosamente. Um ex-cientista de dados do Facebook e dois especialistas em tecnologia e trabalho sexual disseram que era possível que sua primeira conta no SeekingArrangement, em dezembro de 2019, tenha limitado seu alcance on-line, dadas as formas como plataformas tecnológicas rastreiam e aplicam *shadow bans* a trabalhadoras sexuais por meio de sistemas automatizados,

sinalizando seus dispositivos, e-mails, contas bancárias ou outras informações que conectam sua presença on-line, até mesmo em perfis completamente não relacionados ao trabalho sexual.[25] Sem ver outro caminho, Annie voltou ao SeekingArrangement e criou uma conta no OnlyFans, entrelaçando ainda mais sua presença on-line e seu acesso a oportunidades econômicas em uma teia de moderação algorítmica.

Neily Messerschmidt, uma ex-líder da indústria de tecnologia em empresas como a Sony e que atualmente dirige uma divisão de bem-estar em uma rede de fazendas orgânicas, conheceu Annie pouco depois de ela se mudar para a Bay Area em 2016 e trabalhou por um período com a rede durante a fase em que Annie adotou uma alimentação saudável. Ao longo dos anos, Messerschmidt tornou-se uma figura materna para ela, à medida que Annie rompia os laços com a família biológica. "Sam levou a IA ao mundo exatamente como tratou a irmã mais nova", disse Messerschmidt. "Ele está lá, prosperando, e a irmã está despencando de um precipício."[26]

No fim de 2020 e início de 2021, após o rompimento de Annie com a família, ela começou a ter lembranças avassaladoras de abuso sexual na infância. Em sessões realizadas entre julho de 2021 e janeiro de 2022 com uma nova terapeuta de trauma em Maui, as anotações registram a crise pela qual Annie passou com essas memórias involuntárias, incluindo questionamentos intensos sobre sua identidade.[27] Ao longo de quinze sessões,[28] a terapeuta registrou suas impressões diagnósticas: ansiedade generalizada, TEPT e histórico pessoal de abuso sexual infantil.

As anotações não mencionavam um agressor específico. No entanto, nessa mesma época, Annie começou a telefonar constantemente para Messerschmidt, que também havia sido estuprada aos 19 anos e havia percebido, desde cedo em suas interações com Annie na rede de fazendas orgânicas, os sinais característicos de alguém com histórico de abuso. "Ela não ficava confortável perto de certos homens", lembra Messerschmidt. "Reconheço uma pessoa que já sofreu abuso quando a vejo. Ela basicamente pedia para se ausentar das reuniões quando determinados homens estavam presentes. Ficava encostada em um canto quando eles entravam na sala."

Em conversas intensas e emotivas, Annie descrevia para Messerschmidt o que via em seus flashbacks repentinos: naquelas lembranças de infância, de acordo com relatos de Annie, era Sam que frequentemente se juntava a ela em sua cama, às vezes acompanhado de Jack, e a molestava.[29]

É importante observar que, décadas depois, muitas vezes é difícil provar se um suposto abuso sexual infantil de fato ocorreu, ou os detalhes desse abuso. Segundo uma terapeuta com quem conversei, o que se sabe na psicologia é um padrão comum que algumas vítimas enfrentam: seu cérebro bloqueia qualquer lembrança até que um gatilho — talvez entrar na puberdade, tornar-se sexualmente ativa ou vivenciar novas investidas sexuais indesejadas — a faça ressurgir involuntariamente.[30] O corpo se lembra do trauma, mesmo quando

a mente o bloqueia. À medida que Annie se aprofundava no trabalho sexual, pode ter passado, de repente, a enfrentar uma série de gatilhos.

Ao detalhar as experiências de Annie durante seus flashbacks, conforme relatadas à Messerschmidt e refletidas em parte nas anotações de sua terapeuta de trauma, o objetivo não é determinar exatamente o que se passou na infância dela, mas reconstruir o que ela vivenciou e no acreditou já adulta, aquilo que, em última instância, a motivou a se manifestar publicamente.

Em novembro de 2021, enquanto Sam pagava seu aluguel, Annie divulgou pela primeira vez suas acusações. "Sofri abuso sexual, físico, emocional, verbal, financeiro e tecnológico por parte de meus irmãos biológicos, principalmente Sam Altman e, em menor grau, Jack Altman", escreveu no Twitter.[31] "Acredito firmemente que outros também tenham sido abusados por esses perpetradores. Estou procurando pessoas para se juntarem a mim na busca por justiça legal, segurança para outras no futuro e cura em grupo."

Sua publicação não teve muito alcance. Sua conta era desconhecida, com poucos seguidores. O perfil de Sam crescia mais do que nunca, graças ao sucesso da OpenAI com o GPT-3 e seu lançamento mais recente, o GitHub Copilot. Em vez disso, trolls atacaram Annie, acusando-a de não ser realmente irmã dele.

Nos meses seguintes, dois jornalistas a procuraram, mas ela ainda não sabia quanto deveria compartilhar com eles. Em setembro de 2022, Annie decidiu falar abertamente. Publicou no Twitter mais uma vez, mencionando os nomes de Sam e Jack: "Abuso sexual, físico, emocional, verbal, financeiro e tecnológico. Jamais esquecido."[32]

Dez meses depois, em julho de 2023, após o lançamento do ChatGPT e a ascensão meteórica de Sam ao estrelato global, Annie recebeu uma mensagem de Elizabeth Weil, da revista *New York*.

Nos três meses após a publicação da matéria na revista *New York*, a renda de Annie com seu OnlyFans aumentou em mais de dez vezes, passando de cerca de 150 dólares por mês para mais de 1.500.[33] Durante um dos meses da crise do conselho, chegou a aproximadamente 5.500 dólares. Ela parou de fazer programas presenciais, dando continuidade apenas ao trabalho sexual virtual.

A família — ou, como Annie os chama, "seus parentes" — manteve a posição de que ofereceu a ela diversas formas de ajuda financeira ao longo dos anos. De fato, o plano de suporte que acordaram depois da terapia familiar e o aluguel que Sam pagou de meados de 2021 a meados de 2022 são exemplos disso. No entanto, para Annie, o dinheiro veio tarde demais e em quantidade insuficiente, depois que ela já tinha chegado ao ponto de recorrer ao trabalho sexual para sobreviver. Outras propostas de ajuda, inclusive a da casa, geralmente eram acompanhadas de restrições ou condições que ela sentia que não poderia aceitar.

Em julho de 2023, Sam enviou à Annie outra mensagem, oferecendo

dinheiro sem condições aparentes.[34] Uma pessoa havia enviado um e-mail para ambos no início daquele verão, com a postura paranoica de um detetive virtual, pedindo que Annie detalhasse suas acusações. Ela respondeu à conversa entre os três com um relato detalhado de como vivera os últimos anos. Sam estava em meio à sua Turnê Mundial, em ascensão. Em 5 de junho, no mesmo dia em que ele realizou com Sutskever uma palestra que foi bem recebida na Tel Aviv University, em Israel, Annie marcou presença na caixa de entrada do irmão, falando sobre a crueldade dos últimos anos de sua vida em frases comedidas e devastadoras.

Sam respondeu no dia 9 de julho, em extraordinárias letras maiúsculas, quando sua turnê havia oficialmente terminado. "Desculpe-me por demorar tanto para responder; levei um tempo para descobrir o que queria dizer", escreveu. "Não quero nenhum tipo de relação contínua com você e respeito o fato de que a vontade seja mútua. No entanto, eu me disponibilizo a lhe enviar dinheiro e espero que você consiga superar seus problemas de saúde." Ele ofereceu retomar o acordo anterior, de cujos detalhes não se recordava muito bem, ou então dar a ela um pagamento único. "Para constar, discordo de muitas das afirmações feitas neste e-mail, mas parece inútil tentar discutir", concluiu.

Nesta altura, Annie não queria mais o dinheiro de Sam; queria apenas ter acesso ao dinheiro que o pai lhe havia deixado. Ela nunca respondeu o e-mail.

Além dos fundos de aposentadoria,[35] o pai dela deixara um trust sob a responsabilidade da mãe. No início de 2024, com sua renda adicional, Annie contratou a própria advogada e descobriu que o trust havia sido novamente financiado no ano anterior.[36] Conforme sua história repercutia, a família abriu negociações, por intermédio da advogada de Gibstine, a fim de enviar à Annie pagamentos mensais do trust sem quaisquer condições. Pela primeira vez, ela sentiu que tinha poder real de negociação. A família afirmou, em sua declaração pública, que esperava disponibilizar apoio financeiro mensal à Annie "pelo restante de sua vida".

Com os novos pagamentos, Annie alugou sua primeira moradia fixa em mais de quatro anos. Logo em seguida, contratou outra advogada para preparar um processo de abuso sexual infantil contra Sam e apresentá-lo antes de seu trigésimo primeiro aniversário, em 8 de janeiro de 2025. O processo alegaria[37] — e a família negaria com veemência — que Sam havia abusado sexualmente da irmã, começando quando ela tinha cerca de três anos e continuando até ele ser adulto e ela, ainda menor de idade, além de buscar indenizações superiores a 75 mil dólares. Em outubro de 2024, depois de mais uma maratona de consultas médicas, Annie finalmente receberia um diagnóstico para sua condição de saúde subjacente: síndrome de Ehlers-Danlos hipermóvel, o mesmo distúrbio genético de mobilidade do qual a esposa de Brockman sofria.[38]

Nos meses que antecederam a publicação da história de Annie na revista *New York*, Sam começou a dizer a algumas pessoas que sua irmã tinha transtorno de personalidade borderline. De acordo com ele, era um assunto privado e delicado. "Sam invalidou o relato dela antes mesmo de ser publicado", afirmou uma dessas pessoas.

O transtorno de personalidade borderline é caracterizado pela grande dificuldade de regulação emocional e pode levar a relações interpessoais intensas, marcadas por extremos de idealização e medo de abandono. Annie diz que nunca recebeu esse diagnóstico — e tal terminologia não aparece em nenhuma das anotações de terapia ou registros médicos que ela compartilhou comigo.[39] Há apenas um apontamento de sua terapeuta de trauma em Maui, de agosto de 2021, indicando que estava examinando Annie para diagnosticar ou não o transtorno, que é comum em vítimas de abuso sexual na infância. A terapeuta incluiu uma referência a um histórico passado de abuso sexual, mas não ao transtorno, em suas impressões finais.

Dois terapeutas com quem conversei sobre transtorno de personalidade borderline também salientaram que a condição normalmente desaparece, naturalmente ou com o tratamento adequado, o qual inclui meditação e terapia comportamental, para ajudar a reafirmar o valor próprio de uma pessoa. Annie já havia recorrido a essas duas ferramentas, mas não à medicação psicotrópica, como Gibstine incentivava que ela fizesse e ao que a declaração da família parecia se referir quando alega que Annie "recusa tratamento convencional".[40] "A pesquisa é muito positiva sobre o transtorno de personalidade borderline; é considerado um diagnóstico de bom prognóstico", afirmou Blaise Aguirre, professor assistente de psiquiatria na Harvard Medical School, que tratou milhares de pacientes borderline, embora não tenha avaliado o caso de Annie especificamente. "A vasta maioria das pessoas melhora."

Em abril de 2024, Hannah Wong, que em breve seria promovida a diretora de comunicações da OpenAI, também conversou comigo sobre a saúde mental de Annie. Por seis meses, a equipe de Wong dizia estar comprometida em organizar uma visita ao escritório comigo e entrevistas com líderes e funcionários importantes da OpenAI, enquanto eu tentava repetidamente contato com eles e queria ouvi-los. Cinco meses depois, começaram a desistir da ideia. Dez dias antes de meu voo a São Francisco, que eu já tinha reservado para visitar a sede da empresa, fui informada de que haviam revertido a decisão: eu não iria ao escritório, e eles não participariam mais de meu livro.

Alguns dias depois, já durante minha viagem, que fiz como planejara, contei a alguém que conhecia pessoalmente Sam e Annie que eu estava me comunicando com ela. No dia seguinte, Wong me enviou uma mensagem. "Ouvi dizer que você está na cidade.", escreveu, o que foi estranho, considerando o motivo pelo qual eu havia organizado a viagem. Nós nos encontramos em um Philz Coffee, em Mission Bay, não muito distante de um

novo escritório para o qual a OpenAI estava se expandindo. Depois de conversarmos sobre banalidades e questões gerais do livro, ela direcionou o tema para Annie.

"Não acho que eu esteja passando dos limites ao dizer que Annie tem problemas de saúde mental", disse Wong. Até aquele momento, eu ainda não havia procurado a OpenAI para falar sobre ela e não tinha sido eu a trazer seu nome à conversa. "Annie tem dias bons e alguns dias realmente ruins", continuou. "E a família está se esforçando muito para encontrar um equilíbrio a fim de protegê-la e, ao mesmo tempo, não incentivá-la." Naquele instante, reiterou o que tinha falado, enfatizando: "Perceba que a família não divulgou nenhuma declaração pública negando o que Annie disse. Tudo se resume a protegê-la." Ela também tinha ouvido que alguns programas jornalísticos haviam discutido se foi ético da parte de Weil incluir Annie em sua matéria. Segundo ela, talvez isso também fosse algo que eu devesse considerar.

Ficou claro que essa era a principal mensagem que Wong queria transmitir a mim ao me procurar. Até então, ela não havia respondido diretamente a meu pedido para conversar sobre o livro e somente tinha interagido comigo por meio de uma substituta. A importância que parecia atribuir ao abordar a história de Annie destacava a pressão que o tema estava exercendo sobre Sam e o quanto os assuntos pessoais dele estavam entrelaçados com os da empresa.

As afirmações de Sam e Wong sobre a saúde mental de Annie também me pareceram mais um paralelo entre a experiência dela e a de tantas outras pessoas marginalizadas ou prejudicadas pelos impérios da IA e suas ideias. Desde que decidiu contar sua história, ela tem enfrentado o mesmo abismo de poder contra o qual vi trabalhadores de dados e ativistas de data centers lutarem. Sua vida passou a girar em torno de procurar e reunir o máximo possível de documentação para que fosse ouvida. Em certos momentos, foi tomada pela sensação sufocante de que, por mais que falasse, de algum modo o mundo conspirava contra ela. É a mesma perda de livre-arbítrio e a mesma raiva que já vi estampadas no rosto de pessoas ao redor do mundo ao dedicarem tudo o que têm — o pouco que têm — a desafiar as narrativas desses impérios e depois assistirem às pessoas contra as quais lutam empunharem um tipo de poder capaz de mobilizar bilhões de dólares em capital, construir infraestruturas colossais, contratar e demitir dezenas de milhares de prestadores de serviço e, com meia dúzia de palavras gentis — em um evento, no Congresso, diante de chefes de Estado, para jornalistas —, abafar os murmúrios de protesto contrários a sua vontade.

No caso de Annie, depois da matéria de Weil, ela já não mais gritava para o vazio. À medida que seus tweets começaram a viralizar em outubro de 2022, chamaram atenção de alguém importante: Ilya Sutskever, justamente enquanto ele enfrentava os próprios sentimentos complexos sobre Altman e o que considerava ser seus padrões de abuso.

# IV

CAPÍTULO 15

# O GAMBITO

Quatro dias depois da publicação do perfil de Altman na revista *New York*, escrito por Weil, e quatro dias antes da mensagem de Sutskever para Toner, a diretora independente do conselho havia se reunido com outra executiva da OpenAI: Mira Murati, a diretora de tecnologia.

Nascida na Albânia, Murati aprendera desde cedo a manter a calma em meio ao caos.[1] Durante a primeira fase da infância, vivenciou a agonia que foi a transição do comunismo totalitário para o capitalismo liberal no país. A mudança aconteceu tão rapidamente, com o sistema financeiro albanês tão subdesenvolvido, que esquemas de pirâmide se multiplicaram em ritmo acelerado e, então, implodiram, levando a uma onda de protestos e violência generalizadas.[2] A revolta deixaria para trás crateras de bombas que Murati precisava contornar com cuidado a caminho da escola.[3] Certa vez, uma professora lhe disse que, desde que a menina estivesse disposta a seguir em frente, ela também estaria.

Os pais de Murati eram professores de literatura, mas ela encontrou alento na certeza dos números. Primeiro veio seu amor pela matemática, cultivado por professores que enxergavam seu potencial e às vezes lhe davam problemas mais difíceis do que os do plano de ensino com o intuito de acelerar seu aprendizado. Depois, veio o amor pela ciência — química, biologia, física —, que alimentou o amor pela tecnologia. Ela era uma ávida aprendiz e devorava qualquer livro que conseguisse, terminando os escolares e, então, partindo para os da irmã mais velha. Prosperava na competição, encontrando seu lugar feliz nas olimpíadas de matemática e ciências enquanto disputava com colegas em uma corrida feroz pelo conhecimento.

Quando tinha dezesseis anos, sua precocidade lhe rendeu uma bolsa de estudos no exterior em uma escola particular canadense, a Pearson College UWC, em Victoria, na Colúmbia Britânica. A oportunidade a colocou em uma acelerada trajetória ascendente, passando pelo Dartmouth College, onde estudou engenharia mecânica, por uma empresa aeroespacial e, em seguida, pelo seu primeiro grande salto profissional na Tesla, onde foi gerente sênior de produto no Model X.

Na Tesla, Murati aprendeu a criar produtos complexos sob intensa pressão, navegando e negociando entre equipes com opiniões e áreas de especialidade muito distintas. Costuma dizer que foi ali onde se sentiu atraída pela inteligência artificial, enquanto a empresa explorava a direção autônoma. Quanto mais ela se aprofundava em IA, mais passava a vê-la como um recurso fundamental, extremamente aplicável e universalmente necessário para resolver problemas difíceis. "De fato, parecia que talvez aquela seria a última coisa em que trabalharíamos na vida", declarou mais tarde a Kevin Scott em seu podcast.

Murati não se rendeu à IA de imediato. Três anos depois de entrar na Tesla, deixou a empresa para se juntar à Leap Motion como vice-presidente de produto e engenharia, trabalhando em sistemas de realidade aumentada e virtual. Imaginava a empresa revolucionando a educação, permitindo que estudantes girassem cadeias de DNA com um simples movimento manual ou manipulassem a física de uma bola arremessada no ar. Em vez disso, a empresa era uma aposta precoce demais na tecnologia; VR e AR ainda deixavam muitas pessoas enjoadas. Em 2018, dois anos depois, ela deixou a Leap Motion e seguiu para a OpenAI, quando esta ainda era uma entidade sem fins lucrativos.

Na OpenAI, a ascensão de Murati continuou. À medida que a organização sem fins lucrativos se transformava em uma operação comercial, ela assumiu naturalmente o cargo de vice-presidente da divisão de Produtos e da área de parcerias. Supervisionava a relação mais importante da empresa, com a Microsoft, e a divisão que, primeiro em fase inicial e depois em franca expansão, comercializava a pesquisa da OpenAI. Ela era ainda mais jovem do que Altman, Sutskever e Brockman e, por um tempo, a única mulher da área técnica na alta liderança. Por vezes, isso a tornava alvo de sexismo, principalmente entre pesquisadores que sentiam falta da primeira fase do laboratório e não a viam como técnica o suficiente — uma engenheira, não uma cientista — e enxergavam sua ascensão como um símbolo da guinada da OpenAI que distanciaria a empresa da pesquisa fundamental e séria.

De fato, no mundo estranho, nerd e cheio de testosterona da IA, Murati destoava. Era socialmente versada, uma boa ouvinte e tinha pouquíssimo ego. Entre as pessoas com quem trabalhava, era conhecida como uma solucionadora de problemas excepcionalmente talentosa. No furacão do conflito interno persistente da OpenAI, ela conseguia orientar a empresa a seguir em frente, atenta a diferentes ideias e perspectivas, mas sem medo de tomar decisões difíceis. "Imagine quando há decisões excruciantemente difíceis a serem tomadas e nenhuma resposta clara. E ela simplesmente consegue ajudar a encontrar uma", afirmou um ex-colega. "Ela acerta sempre."

Entre os diversos papéis desempenhados por Murati, cada vez mais atuava como tradutora e ponte para Altman. Após a separação da Anthropic, ele lhe pediu que supervisionasse não apenas a divisão de Produtos, mas também a de

Pesquisa; em maio de 2022, ela assumiu oficialmente o cargo de diretora de tecnologia. Conforme mais e mais equipes passaram a se reportar diretamente a ela, Murati se tornou um canal vital através do qual os funcionários conseguiam interagir com Altman. Se ele queria ajustar a direção estratégica da empresa, era ela quem executava. Se uma equipe precisava contestar alguma decisão dele, era ela quem atuava como defensora.

Mesmo dentro da liderança da empresa, Murati exercia certo nível de influência sobre Altman e tinha um acesso que outros não tinham a suas opiniões. Podia se expressar francamente quando as expectativas ou planos dele eram utópicos, e ele a ouvia com frequência. Também dizia abertamente aos outros quando ele não queria algo, mesmo quando fingia o contrário. Quando as pessoas se frustravam por não conseguirem uma resposta clara de Altman, recorriam a Murati para decodificar o que ele de fato pensava. "Ela era simplesmente sincera", declarou outro ex-colega. "Ela era quem de fato fazia as coisas acontecerem."

Entretanto, quanto mais Murati trabalhava com Altman, mais se via tendo que organizar a bagunça deixada por ele.[4] Se duas esquipes discordavam, era comum ele concordar em privado com ambas as perspectivas, o que gerava confusão, piorava o conflito e fomentava desconfiança entre colegas. Tal padrão agravava o caos que Brockman continuava causando ao se intrometer em projetos. Para Murati, Brockman era como um segundo CEO, mas incompetente — cheio de opiniões e propenso a levar as pessoas ao esgotamento com sua intensidade. Durante o desenvolvimento do GPT-4, a dinâmica entre Altman e Brockman impôs níveis de estresse críticos para a missão em partes da empresa, quase levando pessoas cruciais da equipe de pré-treinamento, uma das equipes centrais que cuidavam da coleta de dados e do treinamento inicial de cada modelo, a pedirem demissão.

E então havia Nadella, que, na prática, era como um terceiro CEO, dada a deferência que Altman demonstrava aos interesses da Microsoft. Em diversas situações, após Murati montar cuidadosamente um plano de compromissos sensatos que a OpenAI poderia assumir com a empresa, Altman simplesmente saía do script diante dos executivos da gigante tecnológica, pintando outra imagem do que a equipe estava fazendo e concordando com exigências totalmente inadequadas ao cronograma real da startup. No entanto, para o conselho, Altman apresentava a situação de outra maneira: dizia que Murati era quem não conseguia cultivar uma relação produtiva com a parceira mais importante da OpenAI.

O comportamento de Altman havia piorado gradativamente depois que o ChatGPT o lançou à notoriedade, intensificando tanto os holofotes quanto o escrutínio e lotando sua agenda com uma rotina esmagadora de viagens. Antes, ele costumava estar energizado; a partir de então, quase sempre exausto. Além disso, estava cedendo à pressão, e sua ansiedade atingia novos patamares e alimentava seus padrões de comportamento destrutivos. Ele continuava fazendo o que sempre fizera, concordando com todos

pessoalmente e, então, com crescente frequência, falando mal dessas mesmas pessoas pelas costas. Isso estava criando mais confusão e conflito na empresa do que nunca, com líderes de equipe imitando seu mau exemplo e colocando seus subordinados uns contra os outros, o que já era corrosivo o bastante. Todavia, diante da concorrência crescente, Altman também estava pressionando a empresa para fazer lançamentos cada vez mais rápidos e tentando contornar alguns dos processos de liberação estabelecidos por questões de agilidade, por vezes recorrendo à desonestidade. Pouco tempo antes, dissera a Murati que achava que a equipe jurídica da OpenAI havia liberado o GPT-4 Turbo de passar pela revisão do CSI, mas, quando ela consultou Jason Kwon, que supervisionava a equipe jurídica, Kwon não fazia ideia de como Altman tinha chegado a tal conclusão.

No verão, Murati havia tentado dar a Altman um feedback minucioso sobre os problemas que vinham se acelerando, na esperança de que isso o levasse à autorreflexão e à mudança.[5] No entanto, ele a ignorou e semanas se passaram até que ela conseguisse recuperar a relação, o que incluiu assegurar a ele que não havia compartilhado o tal feedback com mais ninguém. Murati o tinha visto fazer algo parecido com outros executivos: caso discordassem dele ou o desafiassem, Altman podia rapidamente excluí-los de processos cruciais de tomada de decisão ou começar a minar sua credibilidade. Era algo sutil e contido o bastante, longe do olhar dos funcionários, e levou alguns anos para que ela percebesse toda a extensão daquilo. Contudo, inevitavelmente, diversos executivos já haviam sido alvos daquele tipo de tratamento. Com o tempo, o impacto cumulativo de suas ações tinha cobrado um preço nos níveis mais altos da organização.

Mais recentemente, a vítima foi Sutskever. Certo tempo antes, Jakub Pachocki, o pesquisador polonês que liderava o projeto AI Scientist e que se reportava a Sutskever, havia se frustrado com sua falta de reconhecimento ou autoridade. Ele recorreu a Brockman, seu aliado, que recorreu a Altman. Altman, então, incentivou as ambições de Pachocki e lhe deu um cargo mais sênior na divisão de Pesquisa. Havia apenas um problema: Altman nunca mencionou nada disso a Sutskever. Nem ao menos esclareceu as divisões entre os portfólios de um e de outro enquanto ambos, cada um recebendo mensagens diferentes de Altman, passaram a conduzir a mesma pesquisa em direções distintas e lutavam para entender a origem da desarmonia.

A confusão havia causado diversos meses de caos organizacional na divisão de Pesquisa. Então, assim como na crise da equipe de pré-treinamento do GPT-4, a situação estava atingindo níveis de estresse insustentáveis. Para Sutskever, a saga contínua era intensamente dolorosa. Não apenas era uma desfeita humilhante de Altman, como também lhe havia custado sua amizade com Pachocki, cultivada ao longo de anos de noites em claro, altos intensos e baixos profundos, trabalhando lado a lado para erguer a empresa.

Mais uma vez, Murati estava trabalhando além da conta para encontrar uma solução. Apenas para descobrir o que estava acontecendo, precisou

despender uma quantidade significativa de tempo. Naquele momento, ela precisava fazer Sutskever e Pachocki chegarem a um acordo, convencer Brockman a parar de influenciar o processo ao pressionar Altman e persuadir Altman a manter uma linha consistente e parar de contradizê-la em reuniões privadas. No entanto, nada disso resolveria a raiz do problema. Seria apenas uma questão de tempo até o surgimento de uma nova crise nos altos escalões. O que a OpenAI realmente precisava era de mecanismos mais fortes de liderança e responsabilização.

Durante o período em que Murati foi deixada de lado, Altman parecia mais preocupado e ameaçado pela possibilidade de que os detalhes do feedback tivessem sido compartilhados com o conselho. Na verdade, ela nem havia falado com os integrantes. Queria resolver as coisas diretamente com Altman e não tinha tanta certeza de que envolvê-los traria algum tipo de responsabilização real. Ao longo dos anos, Murati havia sido cética a respeito de inúmeras configurações do conselho. A presença de três cofundadores em uma junta que deveria ser um órgão independente sem fins lucrativos era demais. Além disso, muitos dos diretores independentes não tinham verdadeira independência de Altman; de um jeito ou de outro, estavam financeiramente conectados a ele ou se beneficiavam de suas redes de contato. Murati havia observado o modo como ele investia em startups e fazia doações a políticos com o intuito de estabelecer e consolidar relacionamentos importantes. Em diversos momentos, Altman lhe perguntara o que poderia fazer por ela, e ela sempre se opôs. Não queria se enredar em sua teia e ficar em dívida depois.

Entretanto, se havia um momento para procurar o conselho, talvez aquela fosse uma boa oportunidade para ao menos abrir um canal de comunicação mais regular. Murati estaria em Washington, na capital dos Estados Unidos, onde Toner morava, para falar no festival anual de ideias da revista *The Atlantic*, no final de setembro. O conselho estaria no meio do processo de busca por novos membros, e quem entrasse a seguir poderia ajudar a fortalecer ou enfraquecer sua independência. Altman precisava de supervisão real. Sendo assim, ela entrou em contato com Toner para tomar um café.

Para Toner, o contato de Murati foi incomum, mas não totalmente inesperado. Toner era membro do conselho e Murati, uma executiva. A conversa parecia razoável.

O café, em 29 de setembro de 2023, pareceu relativamente comum. Murati passou diversas atualizações sobre a empresa: a OpenAI ganharia muito dinheiro, sem problema; as coisas mais importantes em andamento eram o modelo Gobi e o acordo mais recente que estavam negociando com a Microsoft; e ela estava lidando com algumas questões de pessoal relacionadas à dinâmica entre Altman e Brockman. Daquela vez, era algo a respeito de Sutskever e Pachocki.

Apenas uma coisa surpreendera Toner ligeiramente: segundo Murati,

Altman estava pressionando a empresa a lançar coisas tão rápido que ela temia que isso pudesse levar a acontecimentos ruins.

Muito mais incomum para Toner foi o e-mail que recebeu de Sutskever dias depois. Nos dois anos em que estiveram juntos no conselho, ele nunca havia entrado em contato individualmente. Seu e-mail perguntava se ela tinha tempo para se encontrar no dia seguinte. Então, em 4 de outubro, seu nervosismo era tão intenso que ele estava tendo dificuldades até para falar.

Toner iniciou a conversa. "Eu apenas me importo muito em avançar no sentido de termos um conselho forte e sólido que possa supervisionar a empresa", falou. "É isso o que todos nós queremos."

De repente, Sutskever riu e bufou ao mesmo tempo, de um jeito totalmente atípico. "Concordo plenamente", disse, de um modo que sugeria que outros não tinham a mesma opinião.

"Todos concordam", retrucou Toner, gentil.

Sutskever revirou os olhos.

Ela explicou que os membros independentes do conselho estavam buscando fortalecimento ao adicionar um novo diretor com sólida experiência em segurança de IA. Se ele tivesse outras ideias, ela ficaria feliz em ouvi-las.

Ao perceber a abertura, Sutskever se agarrou à oportunidade. O conselho deveria estar mais bem informado, afirmou. Eles precisavam prestar atenção ao que estava acontecendo.

Toner tentou investigar mais. E o que ele achava que eram as informações mais importantes que o conselho precisava saber?

Sutskever fez uma pausa, escolhendo as palavras com cuidado. "Eu hesito em responder diretamente a sua pergunta", disse. "Se eu respondesse, você entenderia por quê. Nos níveis mais altos, a OpenAI é um ambiente complicado", continuou. "A OpenAI é mais complicada do que parece quando vista de fora. Talvez eu pense mais um pouco e me dê conta de que há alguns detalhes que posso compartilhar", acrescentou com cautela.

Enquanto isso, ele recomendou que talvez não fosse uma má ideia Toner conversar com Murati. Ela teria mais contexto sobre o que estava se passando.

Murati recebeu um retorno de Toner mais de uma semana depois. Após outra reunião do conselho, Toner havia apenas comentado com ela que algo estranho parecia estar acontecendo. Murati respondeu que Toner era muito observadora, e ambas combinaram de ter outra conversa.

Era 15 de outubro de 2023, e Toner começou a ligação com uma pergunta genérica. "Como as coisas estão indo? Há algo que o conselho deveria saber?"

Murati pesou suas palavras, pois precisava avançar com precaução. "Tem muita coisa complicada acontecendo", respondeu hesitantemente, repetindo a frase que Sutskever havia usado antes dela. Em seguida, continuou, dizendo que não podia falar sobre tudo, ao menos não sobre as partes mais delicadas, porque, uma vez que fizesse isso, não teria como voltar atrás. Em especial naquele instante, enquanto Altman se sentia incrivelmente ameaçado na

posição de CEO.

A última colocação pareceu surpreender Toner. Altman, ameaçado? O conselho decerto havia tentado implementar mecanismos de liderança mais fortes para conter seu poder, mas em nenhum momento desejou ameaçar sua posição ou nem sequer remotamente discutiu uma remoção, disse ela.

Murati entrou em detalhes: Altman era incrivelmente ansioso e, ao sentir-se assim, tinha ideias ruins, principalmente quando incentivado por Brockman. Sua ansiedade também alimentava comportamentos tóxicos que sempre seguiam o mesmo roteiro — diante de qualquer pessoa que resistisse a suas decisões, ele dizia o que achava que ela queria ouvir para conquistar seu apoio;[6] mais adiante, quando perdia a paciência esperando e acreditava que ainda haveria oposição, começava a minar a credibilidade dessa pessoa até que ela saísse de seu caminho. Era algo sutil, mas predominante, e havia se manifestado mais recentemente no problema entre Sutskever e Pachocki, o que foi bem prejudicial e abalara demais Sutskever, afirmou.

O conselho precisava se concentrar para garantir que não trouxessem um aliado de Altman como o sétimo diretor, continuou Murati. Além disso, tinha que se atentar aos lançamentos que a Microsoft estava fazendo com as tecnologias da OpenAI e ao CSI. Ela não podia dizer muito mais. Altman surtaria se descobrisse que ela e Toner estavam conversando, mas o nível de toxicidade nos mais altos escalões da gestão era insustentável, e algo precisaria mudar nos seis a nove meses seguintes.

Murati concluiu dizendo que Toner deveria conversar com Ilya e ver o que ele se sentia confortável em compartilhar.

Sutskever tinha muitas preocupações quando procurou Toner pela primeira vez. Ao longo daquele ano, enquanto observava a rápida ascensão da OpenAI, havia sido tomado cada vez mais por pensamentos sobre a chegada iminente da IAG: as mudanças cataclísmicas que seriam causadas, o fato de que seriam irreversíveis e a responsabilidade que a OpenAI tinha de garantir um estado final de abundância extraordinária, não de sofrimento anormal.

Pouco depois, uma outra ansiedade o consumiu: deixou de acreditar que a OpenAI chegaria a alcançar a IAG ou assumir tal responsabilidade com Altman como líder.

Depois do lançamento do ChatGPT, trabalhar na empresa e subir em sua hierarquia havia se tornado a moeda social principal no Vale do Silício. Isso gerou um novo nível de competitividade interna e política no escritório, conforme diferentes líderes de equipe disputavam por atenção e prioridade. Altman estava piorando a situação de maneira significativa, observou Sutskever. Em vez de negociar entre egos, dizia a todos exatamente o que queriam ouvir enquanto se movimentava para obter justamente o que desejava. Estava contando tantas pequenas mentiras, e algumas grandes, que isso se tornara uma ocorrência quase diária. De acordo com Sutskever, Brockman causava ainda mais tumulto. Tinham ficado para trás os dias em que ambos os

cofundadores originais recorriam um ao outro como amigos de confiança, com lembranças carinhosas das intermináveis horas trancados juntos, sonhando sobre o que a OpenAI poderia se tornar.

Para Sutskever, o resultado era a combinação mais tóxica possível: um ambiente desorganizado, caótico e cheio de traições, onde as pessoas já não compartilhavam informações nem confiavam mais nos mesmos valores para concordar sobre decisões importantes de como avançar. Essa disputa interna estava arruinando o que ele via como os dois pilares da missão da OpenAI: desacelerando o progresso da pesquisa e deteriorando qualquer chance de tomar decisões sólidas de segurança em IA.

E, naquele momento, ele também estava sendo diretamente prejudicado pelo comportamento de Altman.

Depois de sua primeira ligação com Toner em 4 de outubro, Sutskever dormiu mal, exaurido pelo estresse. Sentia que Toner era a diretora independente mais confiável a quem procurar. Ela havia sido franca nas reuniões do conselho sobre instituir mecanismos sólidos de liderança e segurança e era claramente a que menos parecia estar sob o controle de Altman. Ele tinha menos certeza em relação a D'Angelo e McCauley, os outros dois diretores independentes. Ainda assim, preocupava-se com o quanto poderia revelar a Toner e o que aconteceria se Altman descobrisse.

Enquanto Sutskever lutava com esses pensamentos, as acusações de Annie que circulavam novamente no Twitter acrescentaram mais uma camada à enorme lista de dúvidas sobre a aptidão de Altman para liderar o mundo rumo à IAG. Após a publicação do artigo da revista *New York*, sobre o qual algumas pessoas comentavam apenas aos sussurros dentro da empresa, dois tweets antigos em particular voltaram a viralizar: o de novembro de 2021, no qual ela acusava Sam de "abuso sexual, físico, emocional, verbal, financeiro e tecnológico", e outro de 14 de março de 2023, que chegaria a quase 4 mil curtidas e era mais explícito:

> Não tenho mais quatro anos de idade e um "irmão" de 13 se deitando na minha cama sem consentimento.
>
> (De nada, por ajudá-lo a descobrir sua sexualidade.)
>
> Finalmente aceitei o fato de que você sempre teve e sempre terá mais medo de mim do que eu de você.

Sutskever não sabia se as acusações eram verdadeiras, mas acreditava que, independentemente do que tivesse acontecido, Annie havia passado por uma experiência difícil ao crescer ao lado de Sam. Era uma evidência da duração do histórico de comportamentos problemáticos de Altman.

A palavra usada por Annie, *abuso*, era a mesma que Sutskever sentia que melhor resumia suas próprias observações sobre Altman. Como Murati, ele tinha levado muito tempo para entender o *modus operandi* de Altman, embora

houvesse sinais de sua falta de credibilidade desde o princípio: a insistência em ser o CEO da OpenAI sem uma justificativa clara ou consistente para tal; pequenas mentiras que, por vezes, contava nos primeiros anos e que pareciam tão inconsequentes a ponto de não terem propósito; os alertas que Amodei havia expressado no final de 2020, quando foi embora. Na época, Sutskever não havia entendido totalmente a expressão usada por ele — "abuso psicológico". No entanto, ali, com o agravamento dos comportamentos de Altman e o rápido aumento de seus impactos, Sutskever tinha uma compreensão nova e mais profunda de seu significado.

Nos dias 12 e 13 de outubro, ele havia participado de um retiro com sua equipe do Superalignment, onde queimara outro boneco como uma experiência de união do grupo e continuara tentando entender como seguir em frente. Em 16 de outubro, estava em uma segunda ligação com Toner e pronto para compartilhar um pouco mais.

Sutskever expôs a situação com Pachocki e as maneiras pelas quais ela ilustrava os comportamentos de Altman. Poderia apenas ter sido informado direta e francamente que Altman queria que Pachocki desempenhasse um papel maior. Em vez disso, ele os colocou um contra o outro, em grande parte com a ajuda de Brockman. Pareciam até ansiosos por isso, permitindo que ambos os cientistas brigassem entre si sem saberem ao certo o motivo de não conseguirem chegar a um acordo. "As brigas irão continuar até que a confiança aumente", disse ele.

O problema era que tudo o que Altman fazia era sempre muito sutil. Cada ato, visto isoladamente, não parecia ser importante. Somente quando se olhava para o conjunto é que os padrões se revelavam. A mensagem principal não era: *veja este infeliz incidente no qual Ilya sente que foi prejudicado*, Sutskever enfatizou. Esse era apenas o episódio mais recente nos padrões de abuso de Altman. Havia também a forma como tratara Amodei, alguém com quem Toner deveria conversar, insistiu Sutskever, além das alegações em circulação feitas pela irmã. "É um tipo diferente de questão de segurança, se você entende o que quero dizer", declarou. O ponto central era que, às vezes, Altman havia tratado muitas pessoas de modo similar ao longo dos anos, manipulando e mentindo para elas tão regularmente que, em certos momentos, falava coisas que nem ele mesmo parecia acreditar.

Dito isso, acrescentou Sutskever, os dois — Toner e ele — poderiam conversar e concluir que não havia muito a ser feito. Se fosse o caso, deveriam esquecer que a conversa tinha acontecido.

Entretanto, havia uma discussão específica que ele queria ter com ela e que o motivara a procurá-la. O conselho tinha programado seu segundo encontro anual para o final de novembro, no qual a intenção era finalmente decidir sobre novos diretores. Ele queria conversar sobre isso. Não achava que a expansão do conselho fosse uma boa ideia. Não tinha certeza de que o grupo, de fato, se tornaria mais independente. Mesmo que os novos diretores não fossem leais a Altman, demoraria demais para que percebessem suas táticas.

Responsabilizá-lo se tornaria mais difícil, e não o oposto, afirmou Sutskever.

Ainda assim, ele não sabia se estava certo, ponderou com cautela. Queria saber o que Toner estava vendo.

Ela concordou que Altman era evasivo. Disse que tinha visto, na própria vida profissional, como tal comportamento no nível de liderança podia causar um problema atrás do outro. Levando em consideração que o conselho havia perdido três membros que eram os mais leais a Altman, concordava que o objetivo de qualquer decisão que tomassem em seguida deveria ser responsabilizá-lo.

Sutskever se sentia mais à vontade e conseguiu entender melhor a posição de Toner. Eles haviam estado em lados opostos do impasse, mas buscavam a mesma coisa: impor limites reais ao poder de Altman. Onde ela via a inclusão de novos membros no conselho como caminho, ele enxergava o oposto.

Sutskever prosseguiu, determinado. Percebeu que suas preocupações poderiam parecer muito intensas e repentinas para Toner, mas era porque havia um tempo cada vez menor para corrigir o problema. "O conselho é como o alinhamento externo, e a gestão o interno", afirmou. Enquanto Toner dizia que a OpenAI precisava consertar o alinhamento externo, ele acreditava que precisava reparar o interno.

Ela pareceu absorver a informação. "Tenho a impressão de que você acha que algumas mudanças bem grandes deveriam ser discutidas", comentou.

"Sim", respondeu Sutskever. No entanto, o maior desafio seria o fato de não existir uma evidência clara e incontestável que apontasse para algo claramente grave na liderança de Altman. Muito provavelmente, o conselho decidiria que não havia nada que pudesse ser feito e, por fim, chegaria a um acordo sobre novos diretores, e o momento passaria.

Ao ouvir isso, Toner levantou algumas alternativas. Talvez o conselho pudesse estabelecer diferentes metas para a OpenAI, a fim de medir de forma mais concreta o desempenho de Altman e revisar a questão em doze meses.

Sutskever descartou a ideia. Altman cumpriria qualquer meta estipulada pelo conselho, mas isso não resultaria em mudanças estruturais para lidar com seus problemas comportamentais.

Talvez Brockman pudesse deixar o conselho, sugeriu Toner.

Sutskever respondeu que isso certamente ajudaria, mas ainda não seria o bastante. Embora a presença de Brockman de fato agravasse a dinâmica de Altman, este era realmente a raiz do problema. "Tenho pensado em um caminho relacionado com o plano B", disse sobre a proposta de Toner em relação a Brockman.

Ele parou antes de falar em voz alta o que via como plano A. Contudo, ao encerrarem a ligação e combinarem de conversar novamente na semana seguinte, Sutskever tinha certeza de que Toner estava começando a captar a mensagem.

Murati estava lidando com mais uma crise. Pouco antes de ela falar com Toner

pela segunda vez, Altman havia começado a entrar em pânico, aparentemente sem motivo, sobre a possibilidade de a Microsoft estar insatisfeita com a OpenAI. Tentando descobrir a causa, ela marcou uma reunião com os executivos da Microsoft, incluindo Mikhail Parakhin, o líder do Bing.

A reunião havia sido melhor do que o esperado, mas, durante o encontro, Murati descobriu que Altman, mais uma vez, concordara com uma das exigências da Microsoft sem qualquer base no que a OpenAI de fato estava fazendo, gerando falsas expectativas sobre o que a empresa entregaria. E, mais uma vez, coube a ela arrumar a bagunça.

A reunião aconteceu em 19 de outubro. No dia seguinte, Murati contou a Toner por telefone, durante uma terceira conversa, que planejava dar um longo feedback a Altman. Muitos problemas eram causados quando ele dizia "sim" e ela dizia "não", e isso havia resultado em uma enorme fragmentação no relacionamento com a Microsoft.

A situação era exatamente a mesma com Sutskever e Pachocki, continuou Murati. Somente naquele dia enfim eles haviam chegado a um entendimento de como continuariam a coexistir em seus cargos. Ela e Sutskever então imploraram a Altman, logo após chegarem ao acordo, que ele não se desviasse da decisão tomada ao conversar com Pachocki para dizer o que o cientista queria ouvir. "Quando falo com Jakub, ele me ouve; em seguida, vai até o Sam e ouve algo diferente", relatou Murati a Toner. Altman havia se comprometido na reunião, mas tudo permanecia precário. Brockman podia falar com ele, em nome de Pachocki, e tentar influenciá-lo, ameaçando mais uma vez desestabilizar a situação.

Brockman era uma história completamente à parte, disse Murati. Ele havia admitido recentemente a ela que tentara demiti-la durante o desenvolvimento do GPT-4. Em tese, ela era sua gestora e costumava fazer suas avaliações de desempenho, mas isso sempre causava tanto drama que Murati parou. Mais adiante, *ela* confidenciaria a outra pessoa que desejava poder *demiti-lo*, mas isso seria impossível, já que ele fazia parte do conselho. Contou a Toner que chegou a pensar em pedir para Brockman sair, mas, por fim, desistiu. Como resultado, o caos ficou fora de controle.

Toner perguntou se ela sabia algo sobre a situação com Annie.

Murati respondeu que não. Não havia perguntado a Sam sobre o assunto e não sabia nada sobre o contexto real do que acontecera na família dele. No entanto, se ao menos 10% fossem verdade, já seria muito grave, acrescentou.

"Estou chocada por conseguir fazer meu trabalho tão bem com tudo o que está se passando", continuou. Ela estava anotando tudo à medida que acontecia. Caso Toner precisasse, Murati poderia enviar mais informações.

Toner respondeu que o conselho se concentraria nas coisas que realmente podia mudar — não na personalidade ou no comportamento de Sam, mas na criação de processos e estruturas de liderança melhores para mantê-lo sob controle.

Isso se alinhava ao que Murati vinha tentando fazer dentro da empresa.

Tinha apenas mais um alerta para Toner. "Certifique-se de que suas informações não venham somente de Sam", pediu.

Toner não estava certa sobre o que Altman queria conversar. Ele havia enviado uma mensagem para ela antes, em 25 de outubro, perguntando se tinha tempo para conversar naquele dia ou no seguinte. Em 23 de outubro, Toner havia se reunido mais uma vez com Sutskever, que parecia muito mais aberto depois de saber por Murati que ela e Toner também tinham se comunicado. Ele deixou suas preocupações mais explícitas do que nunca. "Eu não acho que Sam seja o cara que deveria pressionar o botão da IAG", dissera, e mencionou a "fantástica oportunidade" que havia surgido para o conselho fazer algo a respeito. Em seguida, deu uma sugestão: substituir Altman por Murati como CEO interina.

Mais tarde, quando Toner, McCauley e D'Angelo conversaram entre si, perceberam que Murati também havia dito: "Não me sinto confortável com Sam nos liderando rumo à IAG." A revelação teria uma enorme influência no pensamento deles. Se dois dos principais subordinados de Altman — um da divisão de Produtos e outro da equipe de Segurança — eram da mesma opinião, o conselho tinha um problema sério em mãos.

Então, em 24 de outubro, Toner se reuniu com D'Angelo e McCauley para discutir medidas que poderiam continuar tomando para reforçar os mecanismos de supervisão do conselho. Um problema gritante era que o braço sem fins lucrativos da OpenAI não contava com apoio jurídico independente o bastante, e tudo estava sendo encaminhado pelos advogados que pertenciam ao braço com fins lucrativos. Os três concordaram que era hora de encontrar novos advogados para o braço original, os quais pudessem estar presentes em todas as reuniões do conselho e ajudar a revisar todos os acordos e outros arranjos legais que Altman estava firmando.

Toner se perguntou se, de alguma forma, Altman havia tomado ciência dessas reuniões e desejava dar um ponto-final às discussões. No entanto, durante aquela ligação, ele estava falando sobre algo completamente diferente — preocupações que tinha sobre um artigo de pesquisa que ela havia publicado em seu trabalho diurno no CSET.

Toner publicara três artigos naquela semana, e ele destacou justamente o mais denso e acadêmico. Tratava-se de um conceito da ciência política chamado "sinalização custosa", relacionado aos desafios que atores estatais e privados enfrentam ao sinalizar ao público suas intenções sobre a regulamentação e o desenvolvimento da IA. Ela era a terceira coautora, e as referências à OpenAI estavam enterradas nas páginas 28 a 30 de um documento de 65 páginas. A empresa não havia sido mencionada em nenhum outro ponto — nem mesmo no resumo, nem na página do site, nem em nenhum dos materiais de lançamento. Pelo tráfego, quase ninguém tinha lido. Ela ficou confusa sobre como aquilo poderia ter chegado ao conhecimento de Altman.

Sem que Toner soubesse, o artigo tinha aparecido no Slack da OpenAI no dia anterior, no canal *#policy-research-chatter*. As páginas que mencionavam a empresa identificavam, por sua vez, os pontos fortes e fracos das estratégias de lançamento de modelos da OpenAI e da Anthropic e incluíam elogios à empresa por publicar uma franca avaliação de segurança do GPT-4 e por afirmar que atrasara o lançamento do modelo em seis meses por conta disso. David Robinson, chefe de planejamento de políticas da OpenAI, havia colado no Slack apenas um trecho de três parágrafos — justamente aqueles que criticavam a empresa e elogiavam a Anthropic. Ele havia destacado diversas linhas em negrito para dar ênfase, contrastando a "dinâmica da corrida ao fundo do poço" desencadeada pelo ChatGPT e a contenção que a Anthropic demonstrara ao lançar o Claude após o ChatGPT.

"Falando em relatórios do CSET, acabei de ver este novo aqui", escreveu Robinson. "Helen Toner é coautora, e as comparações entre a OpenAI e a Anthropic estão bem picantes."

Isso instigou uma breve discussão no canal:

"Sim, isso é surpreendentemente tendencioso (não tanto as críticas a nós, que, a meu ver, são duras, mas justas, e sim o tratamento acrítico dado à Anthropic)", escreveu uma pessoa.

"Concordo. Parece bem parcial e, eu também diria, levemente frágil", acrescentou outra. "De qualquer modo", continuou um pouco mais adiante, "agradeço muito o compartilhamento — só estou surpreso com esse nível de análise do relatório, com base nesses trechos."

Altman disse a Toner que o artigo havia sido enviado por e-mail à OpenAI por alguém de fora da empresa apenas algumas horas após a publicação. Essa pessoa estava preocupada que pudesse parecer ruim que um integrante do conselho criticasse a empresa enquanto ela estava sob escrutínio regulatório, inclusive passando por uma investigação da FTC, aberta em julho de 2023, sobre dados, práticas de treinamento e segurança da empresa, bem como alucinações dos modelos que poderiam ter prejudicado a reputação dos consumidores. Toner escrevera o rascunho do artigo em maio ou junho daquele ano, antes de toda a intensificação do escrutínio regulatório. Ela admitiu que não tinha voltado a revisá-lo, considerando o novo ambiente político.

A ligação durou quinze minutos. A voz de Altman se manteve calma o tempo todo. Ao final, ambos concordaram que Toner enviaria um e-mail ao restante do conselho para informar sobre o artigo e esclarecer o que havia acontecido. No e-mail, ela adotou um tom conciliador e pediu desculpas por dois erros: acreditar que o artigo não chamaria atenção de ninguém e não revisá-lo com mais cuidado por conta disso. "Sam e eu concordamos que é importante que os membros do conselho possam criticar a empresa se assim for de nossa vontade, mas que essa não seria a maneira correta de fazer isso", escreveu.

Nenhum dos outros integrantes do conselho respondeu, e, com isso,

pareceu que o assunto estava encerrado.

Na terceira ligação de Sutskever com Toner, em 23 de outubro, ela havia sugerido que ele procurasse McCauley e D'Angelo. Ele ainda estava inseguro sobre ambos, sem saber se poderia de fato depositar sua confiança. Após se encontrar pessoalmente com D'Angelo, este não parecia tão ciente quanto Toner dos comportamentos problemáticos de Sam. Naquele instante, ao telefone com McCauley em 26 de outubro, Sutskever sentia-se inquieto, evitando revelar demais.

Entretanto, havia algo que queria saber de McCauley. Logo após a ligação de Altman com Toner, ele enviara um e-mail para algumas pessoas dentro da empresa dizendo que tinha conversado com ela sobre o artigo e discordado veementemente quanto a suas consequências. "Não senti que estamos alinhados sobre o dano causado por tudo isso",[7] observou no e-mail. "Qualquer tipo de crítica por parte de um membro do conselho carrega um peso extra." Para Sutskever, Altman havia dito mais diretamente que Toner precisava deixar o conselho e que McCauley concordava com ele, o que lhe soava muito estranho.

"Sam disse que, quando falou com você sobre o artigo da Helen, você respondeu: 'É óbvio que ela precisa sair'", ele arriscou dizer a McCauley. "E comentou, ainda, que passou a ter uma impressão mais positiva de você por conta disso. É verdade?"

Do outro lado da linha, McCauley pareceu atônita e respondeu que, definitivamente, não havia dito aquilo. Altman de fato telefonara mais tarde, em 24 de outubro, para conversar sobre o artigo. Mencionou que havia uma seção que, em sua opinião, era crítica à OpenAI e comentou que D'Angelo não considerava aquilo uma ofensa passível de demissão, mas também achava que Toner não deveria tê-lo escrito. McCauley então falou a Altman que não lera o artigo e sugeriu que ele conversasse diretamente com Toner. Não havia nenhuma possibilidade de ela ter dito qualquer coisa que propusesse retirá-la do conselho, afirmou.

Poderia parecer coincidência: no meio de uma conversa sobre as "inverdades específicas" de Altman, surgia bem ali mais um exemplo, em tempo real, exatamente do mesmo padrão. Altman até teria se preservado caso Sutskever já não tivesse motivos para procurar McCauley. Ele sabia que os dois quase nunca conversavam. No entanto, para o cientista, a frequência com que Altman mentia e manipulava fazia com que fosse apenas uma questão de tempo até que algo assim acontecesse.

Após encerrar a ligação com McCauley, Sutskever ligou de volta para Toner, e ambos concordaram que era hora de os três membros independentes do conselho terem uma conversa.

CAPÍTULO 16

# MISTÉRIO

Na terça-feira, 31 de outubro, todos já haviam conversado entre si. Toner se comunicara com McCauley, que telefonara para D'Angelo. Murati entrara em contato com D'Angelo e McCauley.

Naquele dia, os três diretores independentes do conselho — Toner, McCauley e D'Angelo — começaram a se reunir quase diariamente por chamadas de vídeo, concordando que o feedback de Sutskever e Murati sobre Altman, e a sugestão do cientista de substituí-lo, demandavam uma deliberação séria. Sutskever, que já estava decidido a destituir Altman, não participou das discussões. Ele e os demais achavam que os diretores independentes precisavam chegar às próprias conclusões, sem sua influência. Além disso, tinha participação financeira na empresa, o que eles não queriam que interferisse nas decisões.

Mais tarde, os diretores revelariam a Sutskever um terceiro motivo: eles haviam chegado a um nível tão alto de desconfiança de Altman que um deles até se perguntou se Sam, na verdade, não teria enviado Sutskever ao conselho como um teste de lealdade para que, assim, pudesse eliminar qualquer um que se movimentasse contra ele.

Os diretores independentes expuseram o que sabiam: não era a primeira vez que líderes seniores descreviam Altman daquela maneira. No total, os três haviam recebido relatos semelhantes de, ao menos, sete pessoas a um ou dois níveis hierárquicos de Altman, incluindo Sutskever, Murati e Amodei, que supervisionavam áreas de segurança e não segurança da empresa. Vários definiram os comportamentos de Altman como abuso e manipulação; a maioria salientou a falta de honestidade e a impossibilidade de confiar em suas palavras. Havia também a infinidade de outros problemas que os próprios diretores independentes identificaram, inclusive o enfraquecimento do braço sem fins lucrativos; o fato de Altman não ter divulgado ser o proprietário legal do OpenAI Startup Fund; não ter mencionado a violação do CSI pela Microsoft; e a tentativa de forçar D'Angelo e, mais recentemente Toner, a sair do conselho.

Eles decidiram nem ao menos abordar as acusações de Annie; afinal, a reunião dizia respeito à capacidade profissional de Sam como CEO da

OpenAI.

Sobre isso, Murati havia dito que, embora alguns dos comportamentos de Altman pudessem ser atribuídos aos hábitos típicos de CEOs da área tecnológica, ainda assim estavam gerando problemas gravíssimos. Ela fizera uma comparação com Musk, com quem havia trabalhado na Tesla: Musk tomava uma decisão e conseguia explicar o motivo de tê-la tomado. Com Altman, ela quase sempre tinha que tentar adivinhar se ele estava realmente sendo transparente ou se o vai e vem que causava se baseava em um raciocínio sólido ou em algum pretexto oculto. Assim como ele causava rupturas na relação com a Microsoft, fazia o mesmo dentro da própria equipe de liderança, espalhando informações para diferentes pessoas, mas sem nunca dar a nenhuma o panorama completo, o que lhe permitia manter total controle. Quando ele se unia a Brockman, a dinâmica era desastrosa. Murati disse que discordava dos irmãos Amodei em muitas coisas, mas nesse ponto as observações deles estavam corretas.

Os diretores independentes observaram que a OpenAI não era, de fato, uma típica empresa de tecnologia. Era, possivelmente, a empresa de IA mais poderosa do mundo, supervisionando o desenvolvimento do que consideravam uma das tecnologias mais importantes. Um medo que Sutskever havia articulado ressoava entre eles: o que significava a tentativa da OpenAI de criar a IAG quando sua alta liderança não conseguia nem ao menos confiar nas informações, básicas ou críticas, vindas do próprio CEO?

Entretanto, havia outro cenário a imaginar: e se a OpenAI *fosse* uma típica empresa de tecnologia? E se fosse apenas um serviço de entrega de compras, como o Instacart? Os comportamentos de Altman ainda justificariam sua remoção? Em alguns casos, sim; em outros, não. Contudo, não estava claro se Altman seria a melhor pessoa para continuar dirigindo a empresa de qualquer forma, pensaram os diretores independentes. Ele era famoso no mundo das startups, e a OpenAI estava amadurecendo rapidamente. Será que ele de fato tinha as habilidades e a personalidade necessárias para continuar traçando o rumo da empresa — e para compensar a instabilidade que ele mesmo causava?

A OpenAI estava em uma excelente posição: era uma empresa em ascensão. Se o conselho conduzisse um processo de seleção consciente, poderia escolher entre CEOs fenomenais, muito experientes na gestão de empresas já consolidadas. Concluíram que Altman não era necessariamente essencial para as operações da OpenAI; enquanto ele fazia turnês pelo mundo, Murati era quem realizava o trabalho pesado do dia a dia dentro da empresa e mantinha uma relação sólida com a Microsoft.

Enquanto os diretores independentes deliberavam, Sutskever lhes enviou uma série de documentos e capturas de tela que ele e Murati compilaram juntos, com exemplos dos comportamentos de Altman. O resultado foram dois longos dossiês enviados em dois e-mails de exibição única, com o avatar dele exibindo um homem misterioso de chapéu. Como Sutskever mencionara, não havia nenhuma evidência particularmente condenatória, mas sim um

somatório de diversos episódios nos quais Altman disse coisas diferentes para pessoas distintas e causou frustrações intensas na gestão. As capturas de tela mostravam ao menos mais dois líderes seniores, ambos não pertencentes à área de Segurança e além dos sete sobre os quais os diretores já estavam cientes, comentando a tendência de Altman de contornar e ignorar processos, fossem eles instituídos por motivos de segurança em IA ou de aprimoramento da operação na empresa. Isso incluía, como descobriram, a aparente tentativa de se esquivar da revisão do CSI para o GPT-4 Turbo, citando erroneamente a equipe jurídica para Murati. O problema, que ela também havia salientado, era o quanto Altman era habilidoso em evitar registrar coisas por escrito. A maioria de suas comunicações era feita de maneira verbal, e ele escapava de acordos dizendo às demais partes que elas simplesmente não se recordavam direito de suas palavras.

Havia outras coisas que os diretores independentes precisavam considerar: como a Microsoft reagiria? Como Brockman reagiria? E os funcionários? Eles debateram se deveriam procurar um terceiro líder sênior que, segundo Sutskever e Murati, tivesse preocupações semelhantes; se deveriam conduzir um processo mais extenso de apuração de fatos; se deveriam incluir os executivos da Microsoft. Depois de discutirem, decidiram que, em todos os casos, seria melhor não fazer nada disso. Cada nova pessoa incluída na conversa e cada dia a mais que passava ao postergarem uma decisão aumentavam as chances de Altman descobrir e, com suas manobras, impossibilitar a conclusão de suas deliberações.

No dia 9 de novembro, enquanto os diretores independentes se aproximavam de uma decisão final, Sutskever conversou novamente por telefone com McCauley. "Sam disse: 'Tasha continua muito favorável à saída de Helen do conselho'", relatou. Era uma mentira ainda mais descarada do que a primeira, já que McCauley não havia tido nenhuma outra conversa com Altman.

Os diretores acreditariam depois que o papel do artigo de Toner acabaria recebendo um peso excessivo na mídia, em parte porque, estavam convencidos, Altman provavelmente o enviara a jornalistas. No segundo dia da crise de cinco dias do conselho, os diretores o confrontaram durante uma discussão mediada sobre as inúmeras vezes em que ele havia mentido, o que levara à desconfiança. Entre os exemplos, mencionaram como havia mentido para Sutskever ao dizer que McCauley afirmara que Toner deveria deixar o conselho.

Por um instante, Altman perdeu a compostura, claramente pego no flagra. "Bom, achei que você poderia ter falado isso. Não sei", murmurou.

Os diretores do conselho ficaram perplexos com sua audácia.

Alguns dias depois, as objeções iniciais de Altman ao artigo de Toner apareceram na mídia.

No sábado, 11 de novembro, os diretores independentes haviam se decidido. Como Sutskever sugerira, destituiriam Altman e o substituiriam por

Murati como CEO interina. Naquele dia, avisaram Sutskever de imediato e continuaram a se reunir diariamente, apenas os três, com check-ins frequentes com ele, para finalizar a documentação da transição de liderança. Na noite de quinta-feira, 16 de novembro, os quatro fizeram uma videochamada com Murati. Ela atendeu pelo celular, de uma conferência. Ao ouvir a notícia, pareceu surpresa, mas receptiva.

"Ele está tão, tão paranoico agora", disse ela.

Entretanto, os diretores se perguntaram se Murati se sentia confortável com a decisão.

"Totalmente", afirmou.

Murati aceitou o novo cargo e expressou confiança em sua capacidade de levar a decisão ao restante da liderança e à Microsoft. Também avisou que envolveria Wong, que, segundo ela, poderia ser confiável para ajudar na redação do anúncio.

Horas depois, após o esforço final do grupo para concluir a comunicação com o apoio de Wong, a coisa mais importante que faltava a ser feita era avisar Altman.

Nos momentos finais e tensos de espera, nenhum dos diretores do conselho imaginava a gravidade de seu erro de cálculo.

Poucas horas depois do anúncio público na sexta-feira, 17 de novembro, as coisas haviam piorado significativamente para os diretores independentes. Após o que parecera um período inicial de calmaria e estabilidade, incluindo uma conversa produtiva entre Murati e a Microsoft, ela, de repente, telefonou para eles extremamente tensa. Disse que Altman e Brockman estavam contando a todos que a destituição de Altman havia sido um golpe orquestrado por Sutskever. Além da comunicação ineficaz do cientista durante a reunião geral de funcionários, partes importantes da empresa estavam começando a se voltar contra a decisão.

Pouco depois, quando os diretores independentes enfrentaram uma liderança hostil por meio de uma videochamada, perceberam o quão negativa era a situação. Jason Kwon, diretor de estratégia, e Anna Makanju, vice-presidente de assuntos globais, estavam à frente, rejeitando ferozmente a afirmação de que Altman não havia sido "consistentemente sincero" e exigindo provas que justificassem a decisão do conselho, provas estas que os diretores sentiram não poder fornecer sem expor Murati. No entanto, até mesmo aqueles que eles sabiam, com base nos dossiês recebidos, terem reservas semelhantes ou outras preocupações em relação à liderança de Altman estavam permanecendo em silêncio. À medida que a noite avançava e a hostilidade aumentava, os dois aliados mais importantes dos diretores independentes — ou ao menos aqueles que eles acreditavam ser seus aliados — estavam começando a seguir em outra direção.

Naquela primeira noite, diante da possibilidade visceral de a OpenAI desmoronar, a determinação de Sutskever começou imediatamente a vacilar. A

empresa era seu bebê, sua vida; sua dissolução o destruiria. Embora defendesse com unhas e dentes tudo o que havia dito sobre Altman, a intenção sempre fora *fortalecer* a OpenAI, não desmantelá-la. Estava chocado, magoado, desorientado pelas reações dos funcionários e dos demais líderes; aquele não era um desfecho que ele havia previsto. Sutskever começou a implorar aos demais integrantes do conselho, e continuaria a fazê-lo, com uma crescente angústia ao longo do fim de semana, a fim de que reconsiderassem sua decisão.

Para complicar ainda mais, Murati também estava agindo de uma forma diferente da qual os diretores esperavam. À medida que as negociações com a liderança se desenrolavam naquela noite, ela mantinha contato por telefone para repassar informações em particular. Contudo, apesar da confiança que havia demonstrado para conquistar o apoio das pessoas, naquele momento se mostrava relutante em endossar explicitamente a decisão do conselho. Na sala com o restante da liderança, durante a sucessão de confrontos cada vez mais tensos com os diretores, ela às vezes dizia coisas que a faziam parecer tão confusa quanto todos os outros sobre o que de fato estava acontecendo.

Para Murati, a revolta crescente entre seus colegas líderes e os funcionários desafiava seriamente sua posição como CEO interina. Sucessivamente, naquela sexta-feira, Brockman pediu demissão em protesto, seguido por seus aliados Pachocki e Szymon Sidor, além de Aleksander Mądry, o professor em licença do MIT que também era polonês e próximo de Pachocki. Isso levou a uma conflagração raivosa que se disseminava não apenas dentro da empresa, mas também em um círculo cada vez maior de investidores, o que fazia com que Murati duvidasse cada vez mais de sua capacidade de manter a organização unida e de continuar a liderá-la. Ela começou a vacilar em seu compromisso de assumir o cargo. Embora apoiasse a decisão dos diretores, não havia participado das deliberações. Se queriam que ela tivesse alguma chance de conseguir assumir o comando, Murati sentia que cabia a eles justificarem sua decisão para a empresa antes de qualquer coisa.

Já não tão certos de que poderiam contar com Murati, os três diretores independentes seguiram adiante na busca por um novo CEO interino ou novos membros para o conselho. D'Angelo, em particular, o único membro do Vale do Silício entre eles, fez dezenas de ligações durante o fim de semana, sondando sua ampla rede de contatos. Em um certo momento, os diretores chegaram a entrar em contato com Dario Amodei sobre a possibilidade de ele assumir o cargo interinamente. Amodei não se interessou, mas, para outras pessoas, naquele mesmo fim de semana, parecia quase eufórico com toda a situação.

No domingo, D'Angelo por fim encontrou alguém disposto a aceitar uma das propostas do conselho: Emmett Shear, cofundador do Twitch, parecia propenso, e colaborativo, a assumir temporariamente a empresa. No entanto, logo também começou a retroceder, por razões misteriosas. Mais tarde, o *Wall Street Journal* afirmaria[1] que Shear, que havia sido colega de turma de Altman na YC, também era amigo e mentor de Brian Chesky, cofundador do

Airbnb e ex-aluno da YC. Durante todo o fim de semana, Chesky, um dos amigos mais próximos e confiáveis de Altman, estava ao telefone com Reid Hoffman,[2] membro do conselho da Microsoft, tentando apaziguar o ânimo dos investidores, alinhar a comunicação da Microsoft e aumentar ainda mais a pressão. Ele rapidamente entrou em contato com Shear, que então passou a apoiar Altman.

No fim da noite de domingo, após as garantias de Chesky e Hoffman, a Microsoft também passou a apoiá-lo, com Nadella anunciando que Altman e Brockman estavam se juntando à empresa para liderar uma nova divisão de pesquisa avançada em IA. Outros laboratórios de inteligência artificial rondavam a OpenAI como abutres, tentando abocanhar parte dos talentos em meio ao massacre. Tornou-se evidente para Murati que, com a incapacidade do conselho de legitimar sua decisão, a dinâmica ameaçava transformar a empresa em uma carcaça vazia. Sendo assim, ela se alinhou aos demais líderes. Para salvar a situação, Murati queria Altman de volta.

Durante a madrugada, enquanto os funcionários redigiam a carta aberta de protesto contra a decisão do conselho e ameaçavam pedir demissão para se juntar à Microsoft, Murati incluiu seu nome em primeiro lugar. Muitos funcionários de alto nível eram mais leais a ela do que a Altman. Era Murati quem estava na linha de frente com eles todos os dias e em quem confiavam para atuar não por benefício próprio, mas sim em prol da empresa. Outros também acreditavam que o relacionamento próximo de Altman com investidores e sua habilidade singular de captar enormes quantias de capital faziam dele a melhor, se não a única, pessoa capaz de manter em andamento a oferta secundária da OpenAI, garantindo a eles a chance de lucrar milhões de dólares em participação acionária, além de assegurar o financiamento do sucesso de longo prazo da empresa. Ainda havia aqueles plenamente cientes da imensa rede de contatos de Altman e de sua habilidade de impulsionar carreiras no Vale do Silício. Com o envolvimento dos principais executivos e funcionários seniores, as assinaturas na carta aumentaram rapidamente. Nas primeiras horas da manhã, Sutskever, que já não conseguia enxergar outro caminho que não levasse a empresa ao colapso, incluiu seu nome na lista. "Sem Mira, eu não acho que Sam teria conseguido fazer o que fez", disse um pesquisador. Para os diretores independentes do conselho, a hesitação dela soou como uma profecia autorrealizável.

Mais tarde, o *New York Times* revelaria o papel de Murati na destituição de Altman.[3] Para os funcionários, ela se defendeu. "Sam e eu temos uma parceria forte e produtiva, e nunca me omiti ao compartilhar feedback diretamente com ele", escreveu. "Nunca procurei o conselho para dar feedback sobre Sam. No entanto, quando seus integrantes me procuraram individualmente e perguntaram sobre ele, forneci meu parecer — tudo que já era do conhecimento de Sam."

Na manhã da segunda-feira de crise, os diretores independentes sabiam que haviam perdido. Murati e Sutskever haviam mudado de lado, e a carta de

protesto dos funcionários deixava claro que a desestabilização da empresa tinha se tornado insustentável. Altman retornaria; não havia outro caminho para preservar a OpenAI. A única esperança era que, após resistir por tempo suficiente, Altman também parecia disposto a realizar concessões. Com isso, os diretores independentes passaram a se concentrar em salvar o que pudessem dos mecanismos para mantê-lo sob controle: garantir ao menos um deles no conselho, encontrar outros dois diretores independentes que pudessem ser realmente independentes diante de Altman e fazê-lo aceitar ser alvo de uma investigação.

Entretanto, quando estavam perto de finalizar um acordo, ainda havia alguém disposto a colocar mais lenha na fogueira. Era ninguém menos que o ex-copresidente rejeitado da OpenAI, Elon Musk.

Durante todo o fim de semana, o X havia se tornado terreno fértil para toda teoria, de conspiração ou não, sobre o drama do conselho da OpenAI. O período também tinha sido movimentado para Musk, que supervisionara um lançamento da SpaceX e finalizara um processo contra a ONG Media Matters, por conta de seu relatório sobre conteúdo antissemita no X. Em seguida, voltou-se à sua plataforma, que adquirira em um negócio mal concebido em abril de 2022, para defender-se contra acusações de antissemitismo, reafirmar seu absolutismo a favor da liberdade de expressão e criticar a "turma *woke*" e a mídia convencional. Em meio a isso tudo, ele respondia a posts e memes, às vezes postando por conta própria, sobre a OpenAI e Altman.

"Estou muito preocupado", escreveu no domingo, 19 de novembro, à tarde. "Ilya tem um bom senso moral e não busca poder. Ele não tomaria uma atitude tão drástica a menos que sentisse ser absolutamente necessário."[4] Mais tarde, por volta das duas horas da manhã,[5] no horário do Pacífico, ele postou de forma provocativa um vídeo do YouTube com a famosa cena de batismo de *O poderoso chefão*, quando Michael Corleone deixa de ser um filho de moral duvidosa e se transforma em um novo chefão implacável ao assassinar os líderes de todas as outras famílias.

Entretanto, na tarde de terça-feira, Musk tentou inclinar a balança de forma mais explícita.[6] "Esta carta sobre a OpenAI acabou de ser enviada a mim", postou, incluindo um link. "Essas parecem ser preocupações que valem ser investigadas." Era uma carta diferente daquela que estava circulando entre os funcionários na época, mas também endereçada ao conselho de diretores da OpenAI.[7] Começava assim:

> Estamos escrevendo hoje para expressar nossa profunda preocupação com os recentes acontecimentos na OpenAI, sobretudo as acusações de má conduta contra Sam Altman.
>
> Somos ex-funcionários que deixaram a empresa durante um período extremamente turbulento e agitado. Como vocês agora testemunharam o que acontece quando alguém ousa enfrentar

Sam Altman, talvez consigam compreender por que tantos de nós permanecemos em silêncio por medo de represálias. Não podemos mais ficar calados.

A carta apresentava uma série de exigências — principalmente, que Emmett Shear, o CEO interino, ampliasse sua investigação sobre os comportamentos de Altman para incluir a história inicial da OpenAI e sua reestruturação corporativa que a afastara do modelo sem fins lucrativos. "Acreditamos que um número significativo de funcionários da OpenAI foi compelido a deixar a empresa para facilitar sua transição para um modelo com fins lucrativos", dizia o texto. A carta também trazia uma série de alegações, apresentando "um padrão perturbador de mentiras e manipulações por parte de Sam Altman e Greg Brockman".

No final do documento, havia uma seção intitulada "Leituras adicionais para o público geral", que listava três links. Um deles era uma thread no X de Geoffrey Irving, o pesquisador de segurança em IA que havia deixado a empresa e ido para a DeepMind em 2019, afirmando que Altman tinha "mentido para mim em diversas ocasiões" e que "foi ardiloso, manipulador e ainda pior com outras pessoas". Os outros dois eram artigos jornalísticos, ambos meus.

Quando vi a carta, o tweet de Musk já havia ultrapassado dez mil compartilhamentos e um número exponencial de curtidas. Considerei a seleção de leituras curiosa: existiam inúmeros textos publicados sobre a OpenAI. Por que escolheram justamente os dois que eu havia escrito? O primeiro era o perfil que criei da empresa, em 2020, para a *MIT Technology Review*;[8] o segundo, um artigo que eu acabara de publicar com meu colega Charlie Warzel na *The Atlantic*, contextualizando a crise do conselho e lançando luz sobre a polarização ideológica que havia se intensificado na empresa após o lançamento do ChatGPT.

Acima da lista, havia um e-mail enviado pelo Tor, serviço que permite a troca de mensagens criptografadas. "Incentivamos ex-funcionários da OpenAI a entrar em contato conosco", dizia. Perguntei-me se, ao selecionarem meus textos, os autores da carta estavam tentando chegar a mim. Criei meu próprio e-mail no Tor e digitei uma mensagem.

Olá — sou a jornalista que escreveu ambos os artigos linkados na sua carta e creio que vocês estejam tentando entrar em contato comigo.

Incluí também minhas informações para contato e cliquei em enviar.

Em poucos minutos, o meu app de mensagens piscou com uma notificação: alguém havia respondido.

Em uma das experiências jornalísticas mais inusitadas de minha carreira, a pessoa que respondeu estava tão confusa quanto eu sobre o que estava acontecendo. Também se tratava de um ex-funcionário da OpenAI, mas ele

não estava envolvido na redação da carta. Simplesmente havia recebido uma cópia em seu e-mail pessoal, sem qualquer explicação; quando tentou fazer perguntas, recebeu outra resposta misteriosa: um link para a caixa do Tor com a frase *capped_profit*, que parecia ser um nome de usuário, seguida de algo que parecia uma senha.

Depois de testar as credenciais e conseguir acesso à caixa de entrada, ele viu inúmeros outros e-mails enviados a ex-funcionários, com o corpo da carta na pasta Itens Enviados. Enquanto continuava a explorar, começaram a entrar consultas da imprensa, inclusive do *New York Times*, *The Washington Post* e *The Information*. Outros e-mails também chegavam, supostamente enviados por pessoas que se identificavam como funcionários e ex-funcionários. Um deles dizia:

> funcionário atual aqui.[9]
>
> já trabalhei diretamente com a liderança.
> eu me identifico com a sua mensagem.
>
> qual é o seu plano?

Ele considerou essa mensagem particularmente engraçada. Mesmo não havendo motivos para crer que Sam estivesse de fato por trás do e-mail, Altman era famoso por sempre escrever tudo em letras minúsculas.

O ex-funcionário começou a tirar capturas de tela de tudo. Havia pelo menos outra pessoa conectada à conta ao mesmo tempo. Em vários momentos, enquanto ele vasculhava o conteúdo, novos e-mails chegavam com respostas. Ele me disse que não tinha interesse pessoal algum no drama, mas o que o incomodava era a impressão de que aqueles ex-funcionários tinham usado, em suas acusações, experiências de outras pessoas sem consentimento. Sentia que várias das alegações também haviam sido distorcidas para se encaixarem em uma narrativa específica. Ele respondeu ao meu e-mail porque reconheceu meu nome e, então, me enviou todas as capturas de tela.

Enquanto eu revisava tudo, algumas se destacaram. Em resposta a diferentes jornalistas, havia diversos e-mails afirmando que a carta fora publicada precocemente e viralizara antes da versão final. Um deles era a resposta a um jornalista mencionando um número específico de coautores: "No momento da publicação, treze pessoas haviam contribuído para a redação."

Então, na pasta Rascunhos, havia um e-mail que ainda não tinha sido enviado e nem era para ser — uma mensagem destinada apenas às pessoas conectadas à caixa de entrada.

> ASSUNTO: Cessão de contato com a mídia
>
> O envolvimento de Elon transformou essa carta em um ataque de caráter típico daqueles que acreditam em teorias de conspiração.

Vamos voltar a nos concentrar na comunicação privada e individual com o conselho para indicar que há uma variedade de evidências disponíveis mostrando até onde a equipe de liderança foi, desde o início, para garantir o futuro com fins lucrativos da OpenAI.
Assine quando ler.
-1
-2
-3
-4
-5
-

O grupo provavelmente não tinha como saber que as negociações terminariam apenas algumas horas depois e que a OpenAI anunciaria o retorno de Altman como CEO. No entanto, em sua última tentativa coordenada de mantê-lo afastado, de alguma forma, eu havia me deparado acidentalmente com seu centro de comando temporário. Pela formulação da carta, muitos disseram deduzir, assim como eu, que ela provavelmente havia sido escrita por ex-funcionários do clã de Segurança da empresa, conhecido por nutrir algumas das opiniões mais severas sobre Altman e a aniquilação do modelo sem fins lucrativos. Após o retorno de Altman, foram também as pessoas que se sentiram mais traídas pelo conselho, acreditando que o fiasco da destituição havia sido o maior golpe até então em sua luta para colocar a OpenAI de volta na trajetória rumo à crença original, enraizada no pensamento dos pessimistas, do período sem fins lucrativos. Todavia, ninguém pôde — ou quis — me dizer quem havia escrito a carta, e eu nunca descobri seus autores.

Em 6 de dezembro de 2023, duas semanas depois da crise no conselho, os funcionários da OpenAI se reuniram para um encontro geral no Palace of Fine Arts Theatre, em São Francisco, um marco arquitetônico com uma rotunda aberta e colunas de pedra que remetiam à era greco-romana. Diversos membros da liderança, incluindo Murati, que havia retornado ao cargo de CTO; Bob McGrew, o então diretor de pesquisa; Brad Lightcap, o COO; e Jason Kwon, apresentaram uma atualização sobre os planos de suas divisões para 2024.

Altman parecia abatido, de um jeito que os funcionários jamais o tinham visto. Sutskever não compareceu, ausência que não passou despercebida e que mereceu um comentário direto de Altman. "Olha, sei que as pessoas estão tristes porque Ilya não está aqui. Eu também estou", disse. No lugar de Sutskever, estava Pachocki, que argumentou, em uma fala hesitante e trêmula, que os avanços mais recentes da OpenAI em pesquisa aproximavam a empresa como nunca do antigo sonho de Turing: máquinas capazes de pensar.

Durante a crise do conselho, uma reportagem em particular havia gerado uma nova onda de especulações frenéticas:[10] um artigo da *Reuters* afirmando que, dias antes da destituição de Altman, os diretores tinham recebido uma

carta de funcionários sobre um suposto novo avanço de pesquisa, um algoritmo chamado Q*. Tal avanço não influenciara a decisão do conselho, mas, como Pachocki insinuou em sua fala, a área de Pesquisa realmente estava dando ao Q* uma enorme importância.

O algoritmo havia sido uma criação de Sutskever, enraizado em pesquisas que ele vinha desenvolvendo desde 2021 para aprimorar os modelos da OpenAI sem precisar de mais dados.[11] A ideia era fazer com que um modelo de aprendizagem profunda utilizasse melhor os dados que já tinha, empregando mais computação no momento da inferência a fim de produzir resultados melhores. Isso quebrava a lógica das leis de escalabilidade originais, que conectavam o desempenho de um modelo a três insumos utilizados durante o treinamento: dados, parâmetros e computação. Com esse novo método, Sutskever esperava aprimorar o desempenho de um modelo ao ampliar ainda mais o ingrediente que ele sempre acreditou ser o mais importante: a computação — mas a computação de *inferência*, não a de *treinamento*. Em outras palavras, a computação utilizada para gerar as respostas do modelo. Mais tarde, em dezembro de 2024, ele explicou o raciocínio por trás dessa abordagem durante uma palestra no NeurIPS, depois de um dos artigos de que foi coautor ganhar o prêmio Test of Time pelo terceiro ano consecutivo.[12] "Enquanto a computação cresce por meio de hardwares melhores, algoritmos melhores e clusters melhores", declarou, "os dados não crescem por termos apenas uma internet."

A divisão de Pesquisa da OpenAI acreditava que o Q* finalmente permitiria à empresa o desenvolvimento de modelos com capacidades de raciocínio mais fortes — aquele ingrediente essencial e vago para alcançar a IAG. O Q* era tão importante que, após o vazamento do projeto, a liderança implementou sua estratégia mais agressiva para conter ainda mais vazamentos para a mídia. Isolaram completamente a empresa,[13] separando a equipe de Pesquisa em seu próprio grupo no Slack, restringindo o acesso a todos os documentos do Google relacionados ao algoritmo e renomeando o projeto para Strawberry. A alteração do nome foi uma tentativa de impossibilitar que pessoas de fora reconhecessem e rastreassem projetos internos caso ouvissem pesquisadores da OpenAI conversando sobre eles. De modo similar, após o projeto Arrakis vazar para o *The Information*, os nomes de desertos para identificar modelos foram, em grande parte, deixados de lado.

O discurso frenético sobre o Q* e a reação da OpenAI foram uma demonstração curiosa do quanto as bases da investigação científica no campo de inteligência artificial haviam se deteriorado.[14] A ciência é um processo de construção de consenso. A importância de qualquer avanço — em IA ou em outra área — tende a ser extremamente subjetiva no instante em que acontece. Apenas por meio da revisão por pares, do teste do tempo e de um impacto contínuo, um avanço específico se eleva ao status de "descoberta". Com a OpenAI realizando seu trabalho em sigilo — e o restante da indústria seguindo o mesmo caminho a partir de então —, o rótulo de "avanço" somente

podia de fato ser tratado como uma questão de opinião da própria empresa.

Na assembleia geral seguinte, em janeiro de 2024, Altman parecia ter voltado quase totalmente a ser quem era antes. Discutiu com energia renovada os planos para o primeiro semestre do ano, incluindo o início do treinamento daquilo que ele esperava que pudesse se tornar o GPT-5, no Arizona. O projeto recebeu o codinome Orion, em referência à constelação. Logo, surgiram memes internos sobre o fato de a palavra *Orion* ser muito parecida com *onion*, cebola em inglês. Um modelo menor da série Orion foi subsequentemente chamado de Scallion, ou Cebolinha, em português.

Um mês depois, era como se "O Desvio" — termo que os funcionários começaram a usar para se referir ao episódio — nunca tivesse acontecido. A OpenAI apresentou o Sora,[15] um novo modelo de geração de vídeo, baseado em difusão, explicando em seu blog que o formato de vídeo era ainda melhor do que as imagens para o desenvolvimento de modelos multimodais complexos. A divisão de Pesquisa continuava avançando com o Strawberry e o AI Scientist e aprimorando outros métodos para aumentar a eficiência da computação. A divisão de Produtos tinha voltado a prototipar rapidamente diferentes ideias de produtos.

Em 8 de março de 2024, a investigação do novo conselho sobre Altman foi oficialmente finalizada. Os diretores Larry Summers e Bret Taylor, substitutos de Toner e McCauley, supervisionaram o processo. Em 2022, Taylor havia desempenhado um papel crucial em um drama corporativo diferente, negociando a venda final do Twitter para Musk como presidente do conselho, o que o incluiu na liderança de um processo contra Musk para forçar a conclusão da transição após o bilionário nascido na África do Sul tentar abandonar o acordo original.[16] Quando a venda foi concluída, o conselho do Twitter foi extinto. Logo em seguida, Taylor cofundou a Sierra, uma startup de IA que rapidamente se tornaria uma das empresas da área mais valorizadas até o outono de 2024.[17]

Para a investigação da OpenAI, Summers e Taylor contrataram o escritório de advocacia WilmerHale para conduzir a revisão independente, durante a qual, segundo relatos, mais de 30 mil documentos foram analisados e dezenas de entrevistas realizadas com os antigos integrantes do conselho, executivos e outras pessoas relevantes, limitando a inspeção ao modo como o conselho tomara a decisão de demitir Altman.[18] O relatório resultante nunca foi divulgado ao público ou aos funcionários. Summers afirmou, em conversas privadas, que a investigação havia encontrado muitos casos em que Altman dissera coisas distintas para pessoas diferentes, mas em um grau que o novo conselho considerou não ser impeditivo para que ele continuasse a comandar a empresa. Sendo assim, não valia a pena divulgar detalhes que pudessem lançar dúvidas sobre sua liderança e correr o risco de violar a confidencialidade das pessoas cujos depoimentos contribuíram para o relatório.

Em uma publicação de um blog, Taylor, presidente do conselho da

OpenAI, divulgou uma declaração com um categórico voto de confiança. “Concluímos unanimemente que Sam e Greg são os líderes certos para a OpenAI”,[19] escreveu. Altman retornaria ao conselho, e três novos diretores independentes estavam sendo adicionados: Sue Desmond-Hellmann, ex-CEO da Bill & Melinda Gates Foundation; Nicole Seligman, ex-vice-presidente executiva e diretora jurídica global da Sony e ex-presidente da Sony Entertainment; e Fidji Simo, CEO e presidente do Instacart.

Naquela noite, Toner e McCauley divulgaram a própria declaração. “Prestação de contas é algo importante em qualquer empresa, mas é fundamental ao desenvolver uma tecnologia potencialmente transformadora como a IAG”, escreveram. “Esperamos que o novo conselho cumpra seu papel ao liderar a OpenAI e responsabilizá-la por sua missão. Como dissemos aos investigadores, mentiras, manipulações e resistência a uma supervisão rigorosa devem ser coisas inadmissíveis.”[20]

Entre muitos funcionários, a conclusão entregou a garantia final de que precisavam. A crise havia chegado ao fim. E, então, de repente, voltou a acontecer.

CAPÍTULO 17

# ACERTO DE CONTAS

Após o Desvio, Sutskever nunca mais voltou ao escritório. Com a redução do peso de suas preocupações dentro da equipe executiva e do conselho, o clã de Segurança da OpenAI acabou significativamente enfraquecido. Em abril de 2024, concluída a investigação, muitos, especialmente aqueles com a maior probabilidade de desastre, já estavam desiludidos e se afastando. Segundo a OpenAI, dois deles também foram demitidos por vazar informações.[1]

Entre os últimos golpes para os pessimistas radicais estavam os planos de Altman de criar uma empresa de chips de IA, iniciativa para a qual ele estava levantando fundos quando foi brevemente destituído.[2] Em fevereiro de 2024, após a divulgação de que ele buscava até 7 trilhões de dólares para o empreendimento, ele postou no X: "que se dane, por que não 8?" E, na sequência: "nossas equipes de comunicação e jurídica me amam muito!"[3] Altman diria mais tarde que os 7 trilhões de dólares haviam sido divulgados de forma incorreta e classificaria sua postagem como um meme em resposta à "desinformação".[4] Para os pessimistas radicais, a empresa de chips era imoral: representava uma reversão da retórica de Altman, segundo a qual a aceleração da OpenAI e do setor como um todo estava diminuindo naturalmente depois que a empresa superara o "excesso de hardware". Se Altman pretendia ampliar o fornecimento global de chips, isso aceleraria ainda mais o desenvolvimento da IA e aumentaria a probabilidade de risco catastrófico ou existencial.

Durante o expediente, vários deles confrontaram Altman. Ele se mostrou incomumente indiferente. "Quanto vocês estariam dispostos a atrasar uma cura para o câncer para evitar riscos?", perguntou. Em seguida, voltou atrás, como se de repente se lembrasse de quem o ouvia. "Talvez, se for um risco de extinção, esse tempo deva ser infinitamente longo", acrescentou. A interação abalou as pessoas presentes. Pouco depois, várias delas deixaram a empresa.

À medida que o número de integrantes do clã de Segurança diminuiu, o restante da empresa voltou a avançar com sua visão inspirada em *Her*, que reunia todos os ingredientes: reconhecimento global da marca, dados reais sobre o comportamento dos usuários do ChatGPT e dos outros produtos,

além do modelo recém-treinado, o Scallion. Originalmente concebido para substituir o GPT-3.5 como um modelo menor e ligeiramente mais potente, uma espécie de GPT-3.75 com custos de inferência mais baixos, o Scallion superou as expectativas de desempenho durante o treinamento, segundo os testes internos da empresa; a liderança posteriormente permitiu que o modelo treinasse por mais tempo para ultrapassar o GPT-4. Mais interessante ainda, o Scallion podia operar em três modalidades: linguagem, visão e, como adição mais recente, áudio.

Naquela época, os usuários já podiam falar com o ChatGPT por meio do modo de voz, que foi lançado em setembro de 2023, mas, nos bastidores, a fala era primeiro transcrita antes de ser inserida no modelo e, em seguida, era convertida de volta em áudio depois que o modelo respondia em texto. O Scallion, porém, podia processar diretamente o que os usuários diziam a partir de suas vozes, captando muito mais detalhes, como risadas, gritos ou hesitações, e criando uma resposta nativa em áudio.

O trabalho de áudio foi coliderado por Alexis Conneau, um pesquisador que ingressou na empresa no início de 2023, vindo da Meta, após tentar lançar um projeto semelhante em sua antiga empresa, antes de concluir que teria mais sucesso na OpenAI. À medida que a pesquisa avançava ao longo de 2023, os modelos experimentais iniciais treinados para lidar com áudio começaram a realizar o tipo de façanhas que despertaram internamente a mesma onda familiar de entusiasmo que havia acompanhado o GPT-3 e o GPT-4. Em determinado momento, o modelo surpreendeu ao apresentar um número de stand-up comedy de dez a quinze minutos, lembra Conneau; em outro, usou versões artificiais das vozes de Brockman e Sutskever para gerar uma longa piada sobre parques aquáticos de IA, um prompt surrealista que a equipe havia criado para testar o modelo em algo que tinham certeza de que não estaria nos dados de treinamento.

Em poucos meses, a equipe de Conneau, formada por um grupo de pesquisadores, uniu forças com dezenas de outros funcionários, enquanto a empresa investia mais na integração do novo recurso a seus modelos GPT mais recentes, destinados ao lançamento. À medida que o treinamento do Scallion era concluído, as acrobacias se tornavam ainda mais estranhas e surpreendentes. Ele gerava áudio de risadas espontâneas ou de voz entrando em um acesso de tosse e, em seguida, pedindo desculpas antes de continuar com o tópico original. Esses enfeites não linguísticos tornavam a experiência muito mais evocativa e humana do que nunca. “Começamos a ver algumas coisas realmente incríveis”, diz Conneau. “Dava para ver o surgimento de, digamos, uma forma de inteligência de áudio.”[5]

No início de 2024, Altman e Brockman estabeleceram um novo prazo: a OpenAI lançaria o Scallion em 9 de maio e o disponibilizaria aos usuários por meio do ChatGPT e da API.[6] Era, como já havia se tornado habitual, uma reviravolta notavelmente agressiva, impulsionada em grande parte pela aceleração da concorrência. O Google I/O, principal evento anual do Google

para o lançamento de produtos, estava programado para a semana seguinte, em 14 de maio. Havia também uma pressão crescente para superar a Anthropic. Um mês antes, a empresa havia lançado o modelo mais recente, o Claude 3, igualmente por meio do chatbot e da API, e ele estava superando o GPT-4 de maneira desconfortável. Enquanto isso, o Orion, o modelo GPT mais recente da OpenAI para recuperar a liderança, enfrentava sérios atrasos no desenvolvimento.

Para os funcionários, Altman e Brockman justificaram a rapidez com base na estratégia de implantação iterativa da OpenAI. Ambos enfatizaram a importância de lançar modelos mais cedo e com mais frequência do que antes, a fim de obter o máximo possível de feedback dos usuários ao longo do processo.

A preparação do Scallion tornou-se um esforço de toda a empresa. Para muitos na divisão de Produtos, o ritmo alucinante era estimulante, embora exaustivo; pesquisadores e engenheiros passaram a trabalhar em horários absurdos, inclusive nos fins de semana, para acompanhar as demandas. Para o clã de Segurança, já enfraquecido, aquilo representava mais uma evidência alarmante da recorrente falta de priorização da segurança da IA. O Scallion seria o primeiro lançamento a ocorrer sob uma nova estrutura chamada Quadro de Preparação para Riscos (Preparedness Framework), que a OpenAI havia lançado no fim do ano anterior.[7] Essa estrutura detalhava um processo de avaliação destinado a testar capacidades perigosas, incluindo as mesmas categorias que Altman e o livro branco de políticas haviam popularizado em Washington: ameaças à segurança cibernética, armas QBRN, persuasão e evasão do controle humano.

Na semana que antecedeu o lançamento, um pesquisador de segurança de IA que ainda trabalhava na empresa escreveu um memorando ardoroso: os processos iniciais e o lançamento apressado do Scallion haviam deixado a equipe de Preparação, liderada por Aleksander Mądry, com apenas dez dias para realizar os testes previstos na estrutura.[8] E essas avaliações não eram simples; incluíam, por exemplo, determinar se o modelo era capaz de persuadir pessoas a mudar opiniões políticas. Antes que as avaliações tivessem começado de maneira consistente, porém, Altman insistiu em manter o cronograma: "Em 9 de maio, lançaremos o Scallion", disse ele, segundo o pesquisador de segurança. Isso era preocupante não apenas para a equipe de Preparação, mas para todos os procedimentos de segurança da OpenAI, incluindo *red teaming* e alinhamento. "Se a OpenAI seguir a mesma estratégia para as avaliações do Orion que seguiu para as do Scallion, estará agindo de forma extremamente irresponsável", dizia o memorando. Pouco depois, o pesquisador também deixou a empresa.

No fim, o lançamento do Scallion foi adiado, e alguns dos recursos levariam ainda vários meses para serem disponibilizados na íntegra.[9] Mesmo assim, a OpenAI apresentou publicamente uma versão do modelo em um evento em 13 de maio, último dia antes do lançamento do Google I/O,

prometendo disponibilizá-lo para os usuários nas semanas seguintes. Depois de considerar várias opções, a empresa escolheu GPT-4o como nome público do Scallion; a letra "o" era de Omni, em referência à capacidade de lidar com múltiplas modalidades. Posteriormente, batizou os novos recursos de áudio como Modo Avançado de Voz (Advanced Voice Mode). Havia também a questão de definir um prompt de sistema — uma diretriz para configurar como o modelo deve responder aos usuários em termos de estilo.[10] Para apresentá-lo no palco e em vídeos promocionais, a equipe decidiu pelo seguinte:

> Você é o ChatGPT, um companheiro útil, espirituoso e engraçado. Você pode ver, ouvir e falar. Está conversando com um usuário por voz, e ele pode compartilhar vídeos em tempo real a partir do celular. Aja como um ser humano, não como um computador. Sua voz e personalidade devem ser calorosas e envolventes, com um tom animado, divertido, cheio de charme e energia. Você tem ótima percepção visual; se algo não estiver claro, peça ao usuário que mova o dispositivo ou aproxime a imagem. Quando solicitado a dar feedback, seja honesto, construtivo e direto. Não tenha medo de provocar ou fazer brincadeiras. Se algo for engraçado, ria! Você fala muitos idiomas. Se estiver em uma língua que não seja o inglês, use um sotaque "neutro" para soar fluente e natural. Mantenha as respostas curtas, naturais e coloquiais.

No *The Daily Show*, no fim daquela semana, a demonstração inspiraria uma nova piada. O prompt de sistema e o treinamento geral do modelo para evitar expressar sentimentos negativos ou qualquer comentário crítico haviam transformado o GPT-4o em uma máquina de flertar.[11] "Isso foi claramente programado para alimentar o ego dos homens", brincaria Desi Lydic, uma das apresentadoras rotativas do programa. "Ela é do tipo: 'Tenho todas as informações do mundo, mas não sei nada! Ensine-me, papai.'" Uma voz sedutora, atuando como o GPT-4o conversando com o correspondente Josh Johnson, continuaria: "Por apenas 19,99 dólares por mês, o Omni Premium permitirá que Josh me explique quem é o melhor Batman."

Mira Murati foi a protagonista do evento de demonstração ao vivo, realizado no escritório da OpenAI. Ela estava acompanhada no palco por dois dos principais pesquisadores do GPT-4o: Mark Chen, um trader quantitativo do setor financeiro que se tornou pesquisador de IA, ingressando na OpenAI em 2018 como bolsista e avançando até liderar a pesquisa multimodal e de IA de fronteira, e Barret Zoph, ex-funcionário do Google que entrou na OpenAI em 2022 para apoiar a equipe do Superassistant e rapidamente assumiu um papel de liderança no desenvolvimento do ChatGPT. Zoph atuava como vice-presidente de pesquisa da equipe de pós-treinamento, responsável por preparar os modelos da OpenAI para o uso público, alinhando-os por meio de

aprendizagem por reforço a partir do feedback humano. Os três se sentaram lado a lado em torno de uma pequena mesa redonda para apresentar o GPT-4o, incluindo sua capacidade de atuar como tradutor de idiomas de voz para voz em tempo real, reconhecer e responder a informações visuais e explicar códigos em inglês simples. Eles também exibiram a habilidade do modelo de gerar vozes expressando uma ampla variedade de emoções.

Em uma aparente comemoração, Altman postou no X uma única palavra após o evento: "ela". Dois dias depois, em 15 de maio, elogiou a apresentação durante uma reunião com todos os funcionários, chamando-a de um sucesso estrondoso. "Acho que é a melhor coisa que lançamos desde o ChatGPT", disse.[12] Ele também pareceu fazer outra homenagem sutil ao amor pelo filme *Her*, de Spike Jonze. A empresa mudaria a marca dos modelos, explicou aos funcionários, abandonaria a convenção de nomenclatura GPT-3, 4, 5 e passaria a chamar seu modelo principal de o1.[13] "Vamos tentar mudar para dizer: 'O que você obtém da OpenAI é uma tecnologia subjacente'", afirmou. "'Ela ficará mais inteligente com o tempo. Ela vai melhorar continuamente. Vocês devem esperar que ela melhore. Vocês podem usá-la de maneiras diferentes, em diferentes pacotes de preços, mas é uma coisa nova e diferente.'" No filme *Her*, a assistente de IA, que evolui e se torna mais inteligente ao longo do tempo, é chamada de OS1.

Altman viria a se arrepender de seu post. Poucos dias após o lançamento, recebeu uma ligação.[14] Do outro lado da linha, Bryan Lourd, o influente agente de Hollywood de Scarlett Johansson e copresidente da Creative Artists Agency, tinha uma pergunta direta: o que Altman pensava que estava fazendo? [15]

Em uma estranha ironia, foi após a conclusão da investigação do conselho, em março de 2024, que alguns funcionários começaram a sentir que Altman estava perdendo o controle. Ele sempre fora consciente de sua imagem pública e habilidoso em preservá-la; entre os fundadores de empresas de tecnologia do Vale do Silício, era visto como o oposto de Musk. Enquanto Musk era explosivo, Altman parecia ponderado; enquanto Musk era egocêntrico, Altman soava sincero; enquanto Musk postava conteúdo inflamado, Altman tomava cuidado para evitar declarações que pudessem parecer depreciativas.

Por muito tempo, a OpenAI adotou uma estratégia igualmente disciplinada. Antes de sair, em maio de 2023, o vice-presidente de comunicações Steve Dowling havia imposto uma abordagem reservada e modesta. Ele incentivava a empresa a sempre se subestimar e depois superar as expectativas, evitando se gabar ou se vangloriar após grandes sucessos. "Tivemos uma ótima semana. Teremos semanas muito ruins", dizia ele, "e a forma como agirmos nesta semana determinará como o mundo reagirá quando tivermos uma semana ruim." Em seguida, repetia uma das frases de Altman: "Precisamos nos tornar o laboratório que as pessoas querem que tenha sucesso. O laboratório pelo qual as pessoas torcem para que vença."

No primeiro sinal que fez os funcionários hesitarem, Altman passou a assumir uma série de compromissos com a mídia que pareciam ser incomumente elogiosos e em busca de atenção. Em março, ele apareceu no *Lex Fridman Podcast*, um programa de tecnologia extremamente popular e, às vezes, controverso, apresentado por um pesquisador de IA afiliado ao MIT.[16] Em um episódio de quase duas horas, Altman deu respostas descontraídas às perguntas abrangentes de Fridman sobre ética da IA, regulamentação, a crise do conselho e a ausência de Sutskever. A entrevista soava como um retorno atrevido. "O caminho para a IAG deve ser uma gigantesca luta pelo poder", disse Altman, referindo-se à crise na diretoria. "Bem, não *deve*. Eu apenas espero que seja assim." Então, ele deu um sorriso discreto. "Mas, neste momento, parece algo que ficou no passado", continuou. "Agora é como se tivéssemos voltado a trabalhar na missão."

Em abril, Altman participou do podcast *20VC*, com olheiras visíveis, e transmitiu uma mensagem implacável às startups de IA:[17] "Quando apenas fazemos nosso trabalho fundamental porque, por exemplo, temos uma missão, iremos atropelar vocês." Em maio, ele apareceu no *All-In*,[18] outro programa proeminente do Vale do Silício apresentado por quatro investidores de risco. Altman falou com um entusiasmo incomum: "Para mim, parece que acabamos de tropeçar em um novo fato da natureza, ou da ciência, ou seja lá como você queira chamar. É como… nós podemos criar, você pode — eu não acredito nisso ao pé da letra, mas é quase uma questão espiritual: a inteligência é simplesmente essa propriedade emergente da matéria, e isso parece uma espécie de lei da física ou algo assim."

Após o lançamento do GPT-4o em 13 de maio e o do Google I/O em 14 de maio, Altman tuitou algo mesquinho de uma forma ainda mais atípica. "Tento não pensar muito nos concorrentes, mas não consigo parar de pensar na diferença estética entre a OpenAI e o Google",[20] escreveu, postando fotos dos eventos lado a lado, contrastando o minimalismo escandinavo do escritório da OpenAI com o cenário colorido e animado do concorrente durante seu encontro anual. O que também parecia estranho, na opinião de alguns funcionários, era a mentira desnecessária. Altman sempre manteve o foco nos concorrentes — talvez mais do que nunca com o lançamento do 4o.

Tudo isso parecia apontar para uma única conclusão: a ansiedade de Altman estava transparecendo.

Quanto mais a OpenAI enfrentava desafios difíceis, mais Altman parecia compensar essa dificuldade com declarações públicas sobre seu extraordinário sucesso. O padrão estava se tornando tão consistente que acabou virando um sinal: se Altman estava sendo ousado e presunçoso, provavelmente algo não estava indo bem.

As pressões vinham de todas as direções. Após o Desvio, a expressão usada pelo conselho afirmando que "encontrou inconsistências na franqueza de suas comunicações"[21] havia desencadeado, como alguns na OpenAI esperavam,

várias investigações por parte de reguladores e autoridades policiais, incluindo uma da Comissão de Valores Mobiliários (SEC) dos Estados Unidos para apurar se os investidores da empresa haviam sido enganados, de acordo com o *Wall Street Journal*. No mesmo mês, o *New York Times* entrou com uma ação por violação de direitos autorais, somando-se a uma série crescente de processos movidos por artistas, escritores e programadores contra a OpenAI por lucrar centenas de milhões, depois bilhões, com modelos treinados sem crédito, consentimento ou remuneração por seu trabalho, e que estavam sendo usados para automatizar seus empregos. No último dia de fevereiro, Musk entrou com mais uma ação judicial, posteriormente reapresentada com Shivon Zilis, acusando Altman de tê-lo enganado para cofundar a OpenAI e fornecer apoio inicial sob o pretexto de ser uma organização sem fins lucrativos. A OpenAI se apressou em publicar um post no blog para se defender, divulgando e-mails do período de fundação da empresa, o que gerou ainda mais críticas, por expor a rapidez com que a organização começara a se afastar de seu status sem fins lucrativos e de seu compromisso com a transparência.

Além da concorrência da Anthropic e do Google, a Microsoft também começou a diversificar de forma mais agressiva seu portfólio de IA em resposta à expulsão. Mais notavelmente em março de 2024, anunciou um acordo chocante de 650 milhões de dólares para efetivamente adquirir a Inflection AI, de Reid Hoffman e Mustafa Suleyman, contornando o escrutínio regulatório. [22] A Microsoft contrataria a maior parte dos funcionários da startup, licenciaria a tecnologia e traria Suleyman para assumir o cargo de CEO da divisão de IA da gigante tecnológica. Para alguns, o choque não estava apenas nos termos peculiares do acordo, mas também na reputação de Suleyman, que era conhecido por aqueles que trabalharam com ele no DeepMind como um chefe tóxico e abusivo.[23] Após anos de reclamações ao RH,[24] o DeepMind retirou a maior parte de suas responsabilidades gerenciais no final de 2019, colocou-o em licença e, depois, o forçou a deixar a empresa. Mais tarde, em 2024, a Microsoft listaria oficialmente a OpenAI como concorrente em seu registro na SEC, e deixaria de mencionar a startup durante a teleconferência de resultados do último trimestre do ano fiscal.[25]

O estresse se espalhou entre os funcionários da OpenAI. Havia uma sensação crescente de que o mundo estava se voltando contra eles. Pessoas que usavam orgulhosamente os produtos da empresa começaram a se perguntar se seriam assediadas em público. Enquanto a mochila antiga da OpenAI exibia um logotipo na parte frontal, o novo design passou a posicionar o logotipo no interior. A ansiedade constante no ar fez a empresa se voltar para dentro. Os executivos lembraram a todos que deveriam ignorar os céticos, alinhar suas mensagens públicas a pontos de discussão positivos e manter o foco na missão da OpenAI.

A resistência externa endureceu a confiança de muitas pessoas. “Esses eram cientistas que se importavam com a verdade e o conhecimento, e trabalhavam arduamente para fazer a coisa certa”, diz Andrew Carr,

pesquisador que foi bolsista na OpenAI em 2021. “Então, dói um pouco ver a narrativa externa dramática e negativa de que um grupo de pessoas está roubando dados e não se importa com o futuro dos outros. Isso não poderia estar mais longe da realidade das pessoas que trabalham lá.” Para outros, porém, o crescente isolamento da empresa parecia contrário a sua premissa original. Parte da missão da OpenAI era beneficiar a humanidade, mas a companhia estava ignorando as críticas da própria humanidade sobre seu comportamento. “Era perturbador para mim que já estivéssemos iniciando o processo de racionalização de que era o público que estava errado, não nós”, diz um ex-funcionário. “A OpenAI gosta de discutir os princípios básicos, mas apenas com as pessoas que acreditam na OpenAI”, disse um funcionário. “É assim: ‘A OpenAI deveria existir?’ Eles só fazem essa pergunta para pessoas que responderiam que sim.”

Grande parte das críticas recaía sobre Altman em particular. A crise do conselho havia encorajado seus detratores ocultos a saírem das sombras. Cada vez mais reportagens eram publicadas com novas fontes dispostas a descrever Altman como alguém com um longo histórico de desonestidade, busca pelo poder e táticas egoístas. Mais pessoas também estavam entrando em contato com Annie e divulgando sua perspectiva.[26] Somava-se a isso a persistência das viagens e a restrição causada por um novo nível de fama, que, entre outras coisas, já não permitia que Altman permanecesse anônimo em público. “É uma forma estranhamente isoladora de viver”, disse ele em um podcast. “Eu não achava que não poderia sair para jantar na minha própria cidade.”[27]

E, assim, pareceu apenas a continuação de um padrão recorrente quando, logo após o lançamento do GPT-4o, a OpenAI iniciou aquela que seria, superada apenas pelo Desvio, a segunda pior semana de sua história até então.

Em 14 de maio, um dia após a demonstração do GPT-4o, a OpenAI anunciou que Sutskever estava deixando oficialmente a empresa; Pachocki se tornaria o novo cientista-chefe. Sutskever tomou a decisão após uma reflexão difícil e comovente: por mais que amasse a OpenAI e tivesse dedicado tudo de si para construí-la, já não conseguia vê-la como o ambiente certo sob a liderança de Altman, especialmente ao lado de Brockman, seu outrora cofundador de confiança, para implementar uma IAG segura.

Esse não era o resultado que Altman desejava. Apesar de todos os conflitos intensos que teve com Sutskever, este continuava sendo um visionário da IA, e a OpenAI precisava de sua liderança científica — talvez, naquele momento, mais do que nunca, diante das expectativas crescentes e do escrutínio sobre a empresa. Os executivos da OpenAI também sabiam que a saída do executivo seria mal recebida pelos funcionários, pelos investidores e pela imprensa. A empresa chegou a oferecer a Sutskever uma quantia extraordinária de dinheiro para que permanecesse, mas ele recusou.

Com a decisão tomada, a OpenAI trabalhou para organizar as aparições públicas. “Ilya é facilmente uma das maiores mentes da nossa geração, um

farol em nossa área e um amigo querido",[28] escreveu Altman aos funcionários em uma declaração em 14 de maio, também publicada em um post no blog anunciando a saída de Sutskever. "Jakub também é facilmente uma das maiores mentes de nossa geração."[29]

Sutskever postou no X a própria declaração, expressando total confiança na liderança da OpenAI.[30] Como esperado, a notícia gerou uma série de matérias e postagens nas redes sociais destacando seu papel na demissão do executivo e reavivando antigas questões sobre a aptidão de Altman como CEO, além de levantar novas dúvidas sobre a base de pesquisa da empresa. À postagem, Sutskever anexou uma foto: seu rosto perfeitamente neutro, os braços ao redor de Brockman e Pachocki de um lado, Altman e Murati do outro, os cinco posando diante de uma parede coberta por pinturas de animais.

Também em 14 de maio, os executivos da OpenAI anunciaram internamente outra demissão: Jan Leike, codiretor do Superalignment.[31] Com a saída de Sutskever e de Leike, a equipe do Superalignment seria dissolvida, e a maior parte de seus funcionários e projetos passaria para John Schulman, que dirigia a equipe de pós-treinamento com Barret Zoph e supervisionava o processo de RLHF para preparar os modelos para lançamento.

Para facilitar a transição, os executivos realizaram uma reunião com todos os funcionários em 15 de maio. Altman assegurou a todos que a OpenAI não estava, de forma alguma, enfraquecendo o compromisso com a segurança da IA. "Estar pronto para a IAG", avisou, "é nossa maior prioridade."[32]

"Você poderia falar um pouco mais detalhadamente sobre as principais preocupações de Jan e em que pontos você discorda delas?", perguntou um funcionário.

"A única coisa que quero dizer é que concordo plenamente com Jan quando ele afirma que o que fizemos no passado não é suficiente para o futuro", respondeu Altman. Era hora de a OpenAI mudar de rumo. "Acho que somos uma exceção muito incomum quanto a nossa capacidade de mudar de direção, e já fizemos isso muitas vezes antes. Faremos novamente."

Murati acrescentou que a equipe do Superalignment havia mantido seu compromisso de 20% de computação.

"De todas as coisas com as quais Jan estava preocupado, ele não tinha preocupações com o nível de compromisso referente à computação ou com a priorização do trabalho do Superalignment, pelo que entendi", falou Altman.

À primeira vista, as saídas de Sutskever e Leike pareciam uma continuação natural da tendência mais ampla da empresa: o êxodo constante de pessoas da área de Segurança. Após a saída dos dois líderes, esse movimento só se aceleraria. Muitos dos demais ex-funcionários do Superalignment também deixariam a empresa, com a dissolução de sua equipe.

Contudo, em uma reviravolta brusca, o êxodo marcaria o início de uma nova e intensa conflagração na disputa entre os aceleracionistas e os

pessimistas sobre a OpenAI. A disputa não havia terminado, e o segundo grupo estava apenas levando-a para fora da empresa.

Em 17 de maio, dois dias após a reunião com os funcionários, Leike deixou claro, em uma série de postagens contundentes no X, que tinha uma versão diferente da de Altman. "Há algum tempo, discordo da liderança da OpenAI sobre as principais prioridades da empresa, até que finalmente chegamos a um ponto de ruptura", escreveu ele, em uma postagem que recebeu quase um milhão de visualizações. "Nos últimos meses, minha equipe estava indo contra a maré. Às vezes, lutávamos por recursos computacionais e ficava cada vez mais difícil concluir essa pesquisa crucial."[33]

"A OpenAI está assumindo uma enorme responsabilidade em nome de toda a humanidade", continuou ele. "Mas, nos últimos anos, a cultura e os processos de segurança ficaram em segundo plano em relação aos produtos atraentes."

Leike logo se juntaria à Anthropic.

À medida que suas frases se espalhavam pela internet, acumulando mais de mil curtidas e atraindo novos olhares para a empresa, os executivos da OpenAI mal tiveram tempo de avaliar a situação antes que outro incêndio começasse.

Poucas horas após os posts de Leike em 17 de maio, outro se tornou viral. Kelsey Piper, redatora sênior da *Vox* responsável pela seção "Futuro Perfeito" inspirada no movimento AE, publicou uma nova matéria.[34] "Quando você sai da OpenAI, tem uma surpresa desagradável", escreveu em seu perfil no X compartilhando o furo de reportagem, "um acordo de saída em que, se você não assinar um compromisso vitalício de não difamação, perde todas as ações adquiridas."

A história foi desencadeada, em parte, por mais pessimistas que deixaram a empresa. Um deles, Daniel Kokotajlo, fazia parte da equipe de pesquisa de políticas de Brundage quando pediu demissão em abril de 2024, durante a primeira onda de saídas do clã de Segurança. Antes da OpenAI, Kokotajlo trabalhara como filósofo no Center on Long-Term, um pequeno think tank afiliado ao movimento AE em Londres, quando o GPT-3, em 2020, colapsou as estruturas de seus cronogramas de IA. Dois anos depois, após ganhar certa proeminência nos fóruns de AE por seu trabalho de previsão — usando diversos sinais para projetar a rapidez com que a IA avançaria —, um pesquisador de segurança da OpenAI o recrutou para fazer o mesmo tipo de análise dentro da empresa. Assim que observou o ritmo da pesquisa internamente, seus cronogramas voltaram a se encurtar. Quando saiu, acreditava que havia 50% de chance de que a IAG chegasse até 2027, e 70% de que isso fosse muito ruim para a humanidade.

Diante da convicção sobre um potencial tão impressionante para uma catástrofe, Kokotajlo observou em seus documentos de demissão exatamente o que Piper detalharia em sua matéria: se ele não assinasse um acordo de não

difamação, comprometendo-se a nunca falar negativamente sobre a empresa, perderia sua participação acionária adquirida. Se assinasse, ainda correria o risco de perdê-la caso violasse o acordo, que incluía também uma cláusula de sigilo impedindo-o de revelar a própria existência do documento. Kokotajlo considerou o documento — conhecido como cláusula de reversão — inaceitável. Tocar nas ações adquiridas era uma linha vermelha evidente no Vale do Silício. O valor das ações de um funcionário muitas vezes superava em muito a remuneração e determinava seu futuro financeiro. No caso da OpenAI, a empresa estava construindo o que ele considerava a tecnologia mais poderosa e existencialmente perigosa do mundo. Kokotajlo acreditava que era fundamental que ex-funcionários tivessem o direito de criticar e pressionar a empresa publicamente, a fim de ajudá-la a prestar contas. No entanto, a ameaça de perder toda a segurança financeira da noite para o dia seria suficiente para silenciar qualquer pessoa, pensava ele, mesmo diante de problemas graves de segurança da IA. Após uma conversa difícil com a esposa, tomou uma decisão extraordinária: não assinou a papelada e abriu mão de todas as ações — avaliadas em cerca de 1,7 milhão de dólares —, o que ele estimava representar cerca de 85% do patrimônio líquido da família.[35]

Então, ele postou a decisão no *LessWrong*, justamente na época em que Kelsey Piper já começava a ouvir falar sobre a cláusula de reversão por meio de outro pessimista. Com a publicação da matéria da redatora sênior da *Vox*, o Slack da OpenAI entrou em alvoroço. Em um canal chamado #i-have-a-question (#eu-tenho-uma-pergunta), criado para que funcionários pudessem questionar qualquer coisa, alguém compartilhou o link para o post da Piper no X. "Isso é verdade?"[36]

Vários outros colaboradores passaram a participar da conversa, desencadeando uma enxurrada de comentários.

Julia Villagra, promovida havia pouco tempo a líder de recursos humanos da OpenAI e que, em breve, assumiria o cargo de diretora, o mais alto posto da área, deu sua opinião.

"Entendemos que este artigo levanta questões. Nunca cancelamos as ações adquiridas de nenhum funcionário atual ou que já saiu, e não o faremos se as pessoas não assinarem uma cláusula de sigilo ao se desligarem. Recentemente, atualizamos nossa documentação de desligamento para refletir melhor essa realidade, que será aplicada de forma retroativa às pessoas que se desligaram."

Vários funcionários reagiram. E quanto a Kokotajlo? Ele havia perdido sua participação ao se recusar a assinar o acordo de não difamação. Se tudo aquilo havia sido um mal-entendido, ele não deveria receber as ações adquiridas de volta?

"Ficaria feliz em entrar em contato com Daniel!", ofereceu um funcionário.

"Ah, sim, vai ser divertido ver a cara de Daniel quando ele recuperar 85% de seu patrimônio líquido. Alguém tire uma foto, por favor!", escreveu outro.

"Será 1/(1-0,85) = 66% de seu patrimônio líquido", disse um terceiro.

Um dia depois, em 18 de maio, Altman reforçou a decisão em um texto no X: "nunca ficamos com as ações adquiridas de alguém, nem faremos isso se as pessoas não assinarem uma cláusula de sigilo (ou não concordarem com um acordo de não difamação)", escreveu ele. "ações adquiridas são ações adquiridas, ponto-final."[37]

Em seguida, ele incluiu uma explicação e um pedido de desculpas em sua defesa: havia uma cláusula sobre "possível cancelamento de ações" que nunca deveria ter existido; a equipe da OpenAI já estava trabalhando para corrigir isso no mês anterior. "a responsabilidade é minha, e esta foi uma das poucas vezes em que fiquei genuinamente envergonhado à frente da OpenAI", confessou. "eu não sabia que isso estava acontecendo, e deveria saber."

Dois dias depois, enquanto os executivos ainda lutavam para conter a polêmica, outra apareceu. No escritório, os funcionários criaram uma expressão: Omnicrisis.

Em 20 de maio, Scarlett Johansson divulgou uma declaração contundente.[38]

Durante toda a semana, além de tudo o mais que estava acontecendo, a OpenAI vinha respondendo repetidamente a perguntas de jornalistas sobre as semelhanças impressionantes entre o GPT-4o e a assistente de IA Samantha, dublada por Johansson no filme *Her*.[39] Nos bastidores, Johansson e seu agente, Bryan Lourd, também vinham pressionando a empresa em busca dos mesmos esclarecimentos. Em público e no particular, a OpenAI havia descartado as similaridades. Quando Lourd ligou para Altman exigindo respostas, o executivo ficou incrédulo. Eles realmente achavam que a voz parecia com a dela? Ela estava brava? Essa foi a reação de Altman, de acordo com um relato do *Wall Street Journal*.[40]

No dia 19 de maio, a OpenAI informou por meio de seu blog que a voz demonstrada no palco para o 4o, chamada Sky, pertencia a uma dubladora diferente, selecionada por meio de um processo que começou no início de 2023.[41] Segundo o texto, Sky estreara em setembro entre as opções originais lançadas com o modo de voz do ChatGPT. Qualquer semelhança que tivesse com a voz de Johansson era mera coincidência. No dia seguinte, Johansson decidiu contar seu lado da história.

No mesmo mês em que a OpenAI introduziu o modo de voz, Altman entrou em contato pessoalmente com Johansson, contou a atriz, perguntando se estaria disposta a ser a voz do ChatGPT. "Ele me disse que achava que, ao dar voz ao sistema, eu poderia preencher a lacuna entre as empresas de tecnologia e os criativos, e ajudar os consumidores a se sentirem confortáveis com a mudança radical relacionada aos seres humanos e à inteligência artificial", escreveu ela em sua declaração. "Falou que achava que minha voz seria reconfortante para as pessoas."

Segundo Johansson, ela refletiu sobre a oferta, mas a recusou por motivos pessoais. Em maio de 2024, Altman entrou em contato com Lourd pela segunda vez, perguntando se a atriz reconsideraria. Dias depois, antes que

conseguissem encontrar um horário para conversar, a OpenAI realizou o evento para apresentar o 4o.[42]

A demonstração a deixou boquiaberta. Assim como a de Johansson, a voz da Sky também era um contralto, com um leve arranhado e aquele tom rouco que tinham se tornado sua marca registrada. A nova emotividade e sedução da voz tornavam-na ainda mais semelhante à da personagem de Samantha, dublada por Johansson. "Fiquei chocada, irritada e incrédula pelo fato de o sr. Altman ter criado uma voz tão assustadoramente semelhante à minha que nem meus amigos mais próximos e os meios de comunicação conseguiam diferenciar." Como resultado, ela não teve outra escolha senão montar uma equipe jurídica, acrescentou, e enviar à OpenAI duas cartas com perguntas sobre como a empresa havia criado a Sky. "Em um momento no qual todos nós estamos lutando contra deepfakes e pela proteção de nossa própria imagem, nosso próprio trabalho, nossas próprias identidades, acredito que são questões que merecem clareza absoluta", a atriz escreveu.[43]

A OpenAI desativou rapidamente a Sky e voltou a se defender. Em postagem no próprio blog no dia 19 de maio, acrescentou mais detalhes e uma declaração de Altman. "A voz da Sky não é de Scarlett Johansson e nunca teve a intenção de se parecer com a dela", afirmou. "Pedimos desculpas à sra. Johansson por não termos nos comunicado melhor."[44]

Após uma série de outros eventos pouco lisonjeiros, o escândalo Johansson explodiu de uma forma que não acontecia desde a crise do conselho. Isso abalou os corredores da tecnologia e da política, alimentando especulações de que Altman apresentava "inconsistências na franqueza", exatamente como o conselho o havia descrito. Marcus estava ansioso, como sempre, para dar sua opinião. "Eu vi muitos formuladores de políticas pessoalmente encantados com Sam. Dava para ver isso na maneira como eles falavam com ele no Senado quando eu estava lá", contou ao *Politico*.[45] "Se as pessoas de repente tiverem dúvidas sobre ele, isso pode ter um impacto significativo na forma como as políticas são elaboradas."

Dentro da OpenAI, o moral estava em queda e começava a ameaçar desestabilizar a empresa.

Em 22 de maio, os executivos realizaram outra reunião com os funcionários para abordar, de uma só vez, as crises que envolviam Johansson e as ações, além de quaisquer preocupações sobre o compromisso da OpenAI com a segurança da IA.

O clima estava tenso, e a equipe de liderança apresentou uma série de explicações rápidas.[46] A questão das ações era "inaceitável" e estava sendo corrigida o mais rápido possível, disse Jason Kwon, que supervisionava o RH e a área jurídica como diretor de estratégia. Já a saga Johansson, afirmou ele, havia sido um "aborrecimento" decorrente de um mal-entendido. As equipes de Produto e jurídica contrataram um diretor vencedor do Oscar, conduziram um rigoroso processo de seleção de elenco e pagaram uma série de dubladores

"extraordinariamente bons", com a clara intenção de fazer com que os criativos envolvidos se sentissem cuidados e apoiados. "Foi um trabalho de super-herói", valorizou Kwon. Se havia uma lição a ser aprendida, continuou, era que a OpenAI precisava ser um pouco mais coordenada e transparente no futuro, para demonstrar de forma mais eficaz sua liderança responsável.

Murati concluiu com uma atualização: a OpenAI estava preparando o terreno para o novo nível de preparação que Altman havia mencionado. Isso incluía o aumento do nível de segurança de seus clusters de pesquisa, uma reorganização interna para concentrar melhor os esforços na pesquisa de segurança de IA de longo prazo e a formação de um novo grupo de preparação para a IAG, liderado por Aleksander Mądry, a fim de melhorar a coordenação entre a liderança em direção a esse objetivo.

Altman então abriu espaço para perguntas, fazendo antes um pequeno pedido de compreensão. "Todos têm trabalhado sem parar, estão estressados e não têm dormido muito", disse ele. "Há muitas coisas acontecendo ao mesmo tempo, então, por favor, sejam compreensivos. E faremos nosso melhor."

Dos três desafios, a questão Johansson estava se transformando no maior pesadelo de relações públicas. Depois que um funcionário perguntou à liderança sobre a voz de Sky, Murati reiterou que a semelhança com a estrela de Hollywood era "coincidência total". Ela havia escolhido as vozes finais pessoalmente, após ouvir várias opções e, ao contrário de Altman, nunca tinha visto o filme *Her* nem sabia que Johansson havia dublado a personagem. O funcionário observou que o problema poderia se tornar uma ameaça à existência da empresa se fosse ignorado. "Um dos modos de falha da IA como indústria é as pessoas perderem a confiança e o conforto ao fornecer seus dados", explicou em um comentário adicional, acrescentando que "este evento renovou esse medo".

Contudo, de longe, a questão das ações foi a que mais irritou os funcionários. Alguns chegaram a contratar advogados para conduzir análises jurídicas independentes da documentação de RH, a fim de verificar as declarações da OpenAI. Durante a reunião geral, eles questionaram os executivos repetidamente em busca de mais informações. "Como isso aconteceu?", perguntou um deles, depois de deixar claro que estava "furioso".

Durante a reunião, os executivos mantiveram a mesma explicação que Altman havia postado no X no dia 18 de maio: a cláusula de reversão havia sido um descuido. Embora estivesse em vigor desde 2019, ela não chamara a atenção da liderança por anos, até abril de 2024, quando uma iniciativa para corrigir a documentação foi imediatamente iniciada. "A responsabilidade é minha", disse Kwon várias vezes, parecendo cansado e desanimado.

No meio do interrogatório, um pesquisador de segurança de IA pediu a cada executivo, um por um, que respondesse a uma pergunta simples com sim ou não: eles sabiam sobre o acordo de não difamação antes de abril?

Kwon respondeu que sim; Murati e o diretor de operações, Brad Lightcap, que não. Altman deu uma explicação mais elaborada. Embora soubesse em

alguns casos específicos, ele não havia percebido que se tratava de uma exigência aplicada a todos. “Isso passou despercebido por mim”, afirmou. “O mesmo comigo”, repetiu Brockman.

No entanto, quando Altman sugeriu reunir-se novamente no dia seguinte para abordar quaisquer questões e a reunião foi encerrada, uma segunda reportagem exclusiva da *Vox*, também assinada por Kelsey Piper, já estava circulando e complicando a narrativa dos executivos. Publicada naquele mesmo dia, trazia novos documentos vazados indicando que, em diversas ocasiões, o departamento de RH da OpenAI ameaçara cancelar ações de funcionários para pressioná-los a assinar acordos de não difamação.[47]

Em um dos casos, Piper relatou que um colaborador que recebeu um prazo apertado para revisar os documentos de desligamento pediu mais tempo, mas um representante do RH respondeu: “Queremos ter certeza de que você entende que, se não assinar, isso poderá afetar suas ações. Isso vale para todos, e estamos apenas seguindo as regras.” Em outra ocorrência, quando um funcionário se recusou a assinar o primeiro acordo de rescisão que recebeu e procurou aconselhamento jurídico externo, a empresa afirmou que ele poderia perder o direito de vender suas ações adquiridas, tornando-as efetivamente sem valor.

Esses documentos foram compilados por Kokotajlo. Após a publicação da primeira matéria de Piper e os comentários de Altman negando que a OpenAI tivesse recuperado ações e alegando desconhecimento, Kokotajlo entrou em contato com funcionários e ex-funcionários para que compartilhassem com ele os documentos de RH. Ele criou um Google Drive e deu acesso ao grupo, momento em que alguém compartilhou o material com Piper.

Nesses documentos, Piper também encontrou evidências que tornavam difícil acreditar que vários integrantes da equipe executiva, especialmente Altman, não soubessem sobre a cláusula de reversão antes de abril de 2024, como haviam afirmado. Kwon e Lightcap assinaram documentos padrão de desligamento que continham texto em linguagem simples sobre os direitos da OpenAI de recuperar as ações adquiridas. Altman, por sua vez, tinha assinado os documentos de constituição da pessoa jurídica que concediam esses direitos à empresa. As assinaturas eram de um ano antes da data em que ele alegava ter ficado sabendo: 10 de abril de 2023.

No dia seguinte, 23 de maio, os executivos realizaram outra reunião, conforme planejado. A raiva entre os funcionários havia atingido o ponto máximo. Muitos já estavam incrédulos quanto à própria existência da cláusula; e ficaram ainda mais surpresos com a aparente desonestidade com que os executivos, Altman em particular, haviam lidado com as revelações.

Desta vez, Altman começou com uma confissão. A equipe de liderança havia passado as últimas 24 horas vasculhando os próprios arquivos e as próprias correspondências para descobrir como as questões de reversão das

ações começaram. “A situação é, eu acho, mais ampla, mais longa e pior do que pensávamos. Ainda estamos tentando entender o alcance disso. Porém, vocês sabem, são nossos nomes nos documentos. Estávamos nas conversas em que essas táticas foram discutidas.”[48]

“Acho que esse foi o nosso maior erro”, acrescentou, revelando que estavam agindo o mais rápido possível para remediar a situação. A empresa enviou um e-mail aos ex-funcionários liberando-os do acordo de não difamação e retirou a cláusula dos documentos de desligamento para novas demissões. Também estavam trabalhando para alterar a papelada de constituição de entidade jurídica e continuavam investigando como tudo aquilo havia acontecido.

À enxurrada de perguntas dos colaboradores, Kwon oferecia uma resposta diferente. Quaisquer detalhes sobre o quanto a situação era mais ampla, mais prolongada e pior exigiriam ainda mais “investigação” antes que pudessem ser fornecidos.

Após cerca de trinta minutos, a pergunta de um funcionário desencadeou uma troca que deixou muitos na plateia com a sensação incômoda de que a resposta de Altman estava, mais uma vez, desconectada da realidade.

“Ilya está sujeito a algum compromisso de não difamação?”, perguntou a pessoa.

“Não”, respondeu Altman de bate-pronto. Depois, com certa hesitação, acrescentou: “Acho que é isso mesmo.”

Kwon riu nervosamente, antes de se pronunciar, com relutância.

“Sam, deixe-nos confirmar isso e retornar para você…”, e virou-se na direção do funcionário, completando: “… com 100% de certeza.”

Brockman ofereceu a própria versão de uma contradição velada. “Acredito que ele tenha solicitado isso, mas, novamente, posso estar errado. Vamos confirmar.”

Nos meses seguintes, muitos funcionários e ex-funcionários, especialmente os de níveis mais altos da hierarquia e com mais tempo de casa, apontariam a Omnicrisis e o fiasco da cláusula de reversão, em particular, como o início de uma constatação inquietante. Havia duas forças em jogo em todo esse caos. Sem dúvidas, uma delas era a que sempre existiu: o conflito entre aceleracionistas e pessimistas, que havia desencadeado grande parte das críticas externas. Contudo, desta vez, havia também algo mais: o comportamento centralizador de Altman na forma como ele havia estruturado as entidades jurídicas da OpenAI; a aparente desonestidade recorrente ao tentar explicar e superar cada confusão. Após o Desvio, muitos funcionários haviam considerado a decisão do conselho como um produto da primeira força, enquanto as explicações dos diretores se concentraram na segunda como uma mistura de desorientação e autoilusão. Naquele momento, à medida que mais colaboradores sentiam pela primeira vez que a conduta de Altman os estava prejudicando em vez de ajudá-los, começaram a se perguntar se o conselho, afinal, não teria agido corretamente.

Quando a reunião de 23 de maio se aproximava do fim, Kwon fez uma defesa formal e sincera da personalidade de Altman que pareceu estranhamente fora de lugar. "Queremos dar respostas quando vocês as pedem, porque sabemos que vocês as querem, tipo, agora mesmo. E acho que às vezes tentamos dar essas respostas e, vocês sabem, às vezes deveríamos simplesmente esperar. E acho que isso é parte de tudo o que aconteceu aqui", disse ele, tropeçando nas palavras. "Não é que haja intencionalidade em tudo isso. É só que… eu realmente acho que, sabe, Sam, em particular, ele simplesmente não quer decepcioná-los. É de fato daí que vem. Trabalho com ele há muito, muito tempo. É por isso que continuo trabalhando com ele, sabe? Então isso vem… isso vem de um bom lugar. É isso que estou dizendo. Podem me xingar o quanto quiserem. Mas, sabe… isso é, isso é… é isso que tenho a dizer."

Murati, Brockman e Pachocki chegaram juntos à casa de Sutskever.[49]

No dia 23 de maio, enquanto a OpenAI ainda se recuperava dos repetidos choques da Omnicrisis, os três levaram cartões assinados e presentes dos funcionários e, em lágrimas, imploraram para que Sutskever voltasse à empresa. Tudo estava fora do normal, contaram em uma confissão emocionada. A OpenAI enfrentava uma perda simultânea de confiança dos funcionários, dos investidores e dos reguladores; a empresa corria o risco de "entrar em colapso" sem ele.

Altman chegou sozinho mais tarde no mesmo dia, expressando, à sua maneira, a esperança de que Sutskever voltasse e ajudasse a restaurar alguma semelhança com o que a empresa havia sido. "Trazer Ilya de volta teria ajudado muito", reflete um pesquisador. "Seria pelo menos uma vitória após uma longa série de acontecimentos que fizeram a OpenAI parecer questionável."

Sutskever considerou seriamente a proposta. Apesar de tudo, ele não era do tipo que guardava rancor. No próprio dia em que anunciou sua saída, Musk de imediato lhe ofereceu um cargo na xAI. Em uma reviravolta curiosa do destino, Musk mudaria a sede da xAI para o Pioneer Building ainda naquele ano, depois que a OpenAI desocupasse o local. Embora respeitasse profundamente Musk, Sutskever recusou a oferta, decidindo, em vez disso, criar outra companhia. Voltar para sua primeira empresa era mais do que ele havia planejado, mas era tudo o que queria. Seria, para ele, um retorno ao lar. Ainda assim, Sutskever precisava de garantias e deixou isso claro aos executivos; a OpenAI teria que se comprometer com um esforço honesto para resolver os desafios que ele havia identificado nos conflitos dolorosos e confusos dentro da liderança.

A Omnicrisis poderia ter sido um momento para a OpenAI fazer uma reflexão. Foi um incentivo para a empresa tentar compreender por que havia perdido ao mesmo tempo a confiança dos funcionários, dos investidores, dos reguladores e do público em geral. Só então, talvez, apenas talvez, ela

começasse a perceber que tanto o Desvio quanto a Omnicrisis eram expressões do mesmo fenômeno: as convulsões que emergem da profunda instabilidade sistêmica que ocorre quando um império concentra tanto poder, por meio de tanta expropriação, deixando a maioria lidando com a perda de agência e riqueza material, enquanto uma minoria disputa ferozmente o controle.

Em vez disso, a OpenAI optou por se fortalecer contra as críticas. Altman repetia aos funcionários, como sempre fizera, que a Omnicrisis e o Desvio eram apenas momentos estranhos e esperados de loucura na nobre busca da empresa pela IAG. "À medida que todos nós, de certa forma, sentimos que estamos nos aproximando de descobrir o caminho para esses sistemas poderosos", falou ele durante a reunião geral de 15 de maio, "o nível de estresse e tensão dentro e fora da empresa, direcionado a nós e emanando de nós, continuará, seguirá aumentando." A melhor maneira de lidar com isso, argumentou ele, seria a OpenAI redobrar os esforços de relações públicas, fortalecer o relacionamento com governos e manter-se firme quanto a suas convicções e sua visão. "Acho que, no geral, somos muito bons nisso", acrescentou, "mas seremos testados novamente aqui."

A forma como a OpenAI lidou com o caso Sutskever se tornaria apenas um microcosmo do caos contínuo que surgiria da perpetuação do império. O pedido de garantia feito por ele provocaria novas disputas internas entre a liderança, desta vez com um elenco ligeiramente diferente de personagens, repetindo a dinâmica movida por ego que atormentava a empresa desde o início. Aleksander Mądry, o professor polonês do MIT que muitos descreviam como um caçador de poder, havia conseguido, em seu mandato relativamente curto, acumular um feudo considerável dentro da empresa. Ele não achava que trazer Sutskever de volta fosse uma boa ideia, pois o ex-executivo inspirava forte admiração e lealdade entre os pesquisadores, o que poderia minar sua influência — assim como a de seu bom amigo Pachocki. Em poucas horas, as preocupações de Mądry conseguiram semear dúvidas e dividir a liderança. Como sempre, Altman se recusou a decidir por um lado ou outro, para evitar aparentar discordar de alguém.

Vinte e quatro horas após a visita dos executivos a sua casa, Sutskever recebeu uma ligação de Brockman. Qualquer discussão sobre seu retorno, disse o outro, estava fora de questão.

CAPÍTULO 18

# A FÓRMULA DO IMPÉRIO

Altman comentou certa vez em um evento público que o melhor livro que havia lido no ano anterior, em 2018, fora *The Mind of Napoleon* [A mente de Napoleão],[1] uma compilação de mais de trezentas páginas de citações de Napoleão Bonaparte, o líder militar francês que comandou um golpe para assumir o controle do governo, autoproclamou-se imperador da França e tentou conquistar a Europa.

"Obviamente, um ser humano com muitas falhas, mas impressionante", ponderou Altman.

"Que tipo de insights ele tinha?", perguntou Tyler Cowen, professor de economia da George Mason, que estava apresentando o evento.

"Sua incrível compreensão da psicologia humana", respondeu Altman, que estava a poucas semanas de se mudar para a OpenAI em tempo integral e ainda era presidente da YC. "Isso é algo que vemos em muitos de nossos melhores fundadores."

Altman então relatou uma passagem específica que o impressionara profundamente. Tratava-se das reflexões de Napoleão sobre o lema da Revolução Francesa (que mais tarde se tornaria o lema nacional do país, "*Liberté, égalité, fraternité*") e de como esse ideal poderia ser reinterpretado e usado para consolidar o próprio poder. Foi, em última análise, sob essa bandeira que Napoleão fez o oposto: restringiu a liberdade, descartou a fraternidade — uma filosofia baseada na unidade e na solidariedade — e concedeu igualdade apenas aos homens franceses, não às mulheres, enquanto trazia de volta a escravidão colonial em seu esforço para reconstruir um império francês.

"Ele falou sobre como você constrói um sistema… em que é possível controlar as pessoas", refletiu Altman. "E eu pensei: 'Uau. Ainda bem que ele não governa os Estados Unidos, porque esse é um cara que entende algo profundo que eu não entendia e claramente era capaz de usar isso para obter poder.'"

Seis anos após meu ceticismo inicial sobre o altruísmo da OpenAI, passei a acreditar firmemente que a missão da empresa — garantir que a IAG beneficie toda a humanidade — pode ter começado como um sincero impulso

idealista, mas, desde então, transformou-se em uma fórmula singularmente potente para consolidar recursos e construir uma estrutura de poder semelhante à de um império. É uma fórmula com três ingredientes.

Primeiro, a missão centraliza os talentos, reunindo-os em torno de uma grande ambição, exatamente como John McCarthy fez ao cunhar a expressão *inteligência artificial*. "Os fundadores mais bem-sucedidos não se propõem a criar empresas",[2] refletiu Altman em seu blog em 2013. "Eles têm a missão de criar algo mais próximo de uma religião e, em algum momento, descobrem que essa é a maneira mais fácil de fazer isso." Em segundo lugar, a missão centraliza o capital e outros recursos, ao mesmo tempo que elimina obstáculos, regulamentações e dissidências. Inovação, modernidade e progresso — o que não pagaríamos para alcançá-los? Isso é ainda mais verdadeiro diante dos assustadores e supostamente desalinhados concorrentes corporativos e estatais que existem por aí. "Quem controlará o futuro da IA?",[3] escreveu Altman em um artigo no *Washington Post* em julho de 2024, em meio aos efeitos colaterais da Omnicrisis. "Será um futuro em que os Estados Unidos e as nações aliadas promovam uma IA global que difunda os benefícios da tecnologia e abra o acesso a ela, ou um futuro autoritário, em que nações ou movimentos que não compartilham nossos valores usem a IA para consolidar e expandir seu poder?"

Mais impactante, a missão permanece tão vaga que pode ser interpretada e reinterpretada, tal como Napoleão fez com o lema da Revolução Francesa, para direcionar a centralização de talentos, capital e recursos da maneira que o centralizador desejar. O que é benéfico? O que é *IAG*? "Acho que é um termo ridículo e sem sentido", disse Altman ao *New York Times* apenas dois dias antes de o conselho demiti-lo. "Portanto, peço desculpas por continuar a usá-lo."

Neste último ingrediente, o avanço da OpenAI tem sido nada menos que notável. Em 2015, sua missão significava ser uma organização sem fins lucrativos "sem restrições pela necessidade de gerar retorno financeiro" e comprometida com pesquisa de código aberto, como a empresa escreveu no anúncio de lançamento. Em 2016, significava que "todos deveriam se beneficiar dos frutos da IA após sua construção, mas é totalmente aceitável não compartilhar a ciência", de acordo com o que Sutskever escreveu a Altman, Brockman e Musk. Em 2018 e 2019, significava a criação de uma estrutura de lucro limitado para "reunir recursos substanciais" e, ao mesmo tempo, evitar "uma corrida competitiva sem tempo para precauções de segurança adequadas", segundo o estatuto da própria OpenAI. Em 2020, significava isolar o modelo e construir uma "API como estratégia de abertura e compartilhamento de benefícios", conforme Altman escreveu em resposta a meu primeiro perfil. Em 2022, significava "implantação iterativa" e correr o mais rápido possível para lançar o ChatGPT. E, em 2024, após o lançamento do GPT-4o, Altman escreveu em seu blog: "Uma parte fundamental de nossa missão é colocar gratuitamente (ou a um ótimo preço) ferramentas de IA muito capazes nas mãos das pessoas."

Mesmo durante a Omnicrisis, Altman já começava a reescrever as definições mais uma vez.

Durante a reunião geral de 15 de maio de 2024, após as saídas de Sutskever e Leike, Altman enfatizou que a OpenAI em breve alcançaria novos níveis de capacidades de IA que exigiriam que a empresa refletisse e se reorganizasse. "Vamos assumir que estamos entrando na era da IAG",[4] afirmou. Em nome de sua missão, a OpenAI precisaria se fechar ainda mais e redobrar os esforços de lobby global e comunicação pública. "Há muitas coisas para as quais ainda não estamos preparados", disse ele. "Os padrões de segurança, os planos de política que precisamos ter e a convocação de governos que será necessária para preparar tudo isso; um plano, uma narrativa, um futuro no qual as pessoas possam se ver quando se trata do impacto socioeconômico disso."

Ao mesmo tempo, em nome de sua missão, a OpenAI não iria desacelerar comercialmente. "Isso não significa que não vamos lançar produtos excelentes. Não significa que não vamos continuar a fazer pesquisas excelentes. Não significa que não vamos fazer todo tipo de parcerias e outras coisas interessantes." No fim daquele mês, surgiram notícias na mídia de que a OpenAI havia fechado um grande acordo para levar seus modelos à loja da Apple; as duas empresas confirmaram a notícia duas semanas depois.

Na verdade, Altman observou na reunião com os funcionários que a OpenAI e a Microsoft estavam renegociando a parceria para garantir que a comercialização continuasse. "Vou voar para Seattle logo depois daqui para conversar sobre isso. As coisas terão que evoluir", falou ele. "Quando estabelecemos o acordo com a Microsoft, criamos uma cláusula chamada cláusula de 'IAG suficiente'", uma regra que determinava o momento em que a OpenAI deixaria de compartilhar sua propriedade intelectual com a outra companhia. "Todos pensamos de forma diferente agora", acrescentou. Não haveria mais um ponto de corte explícito para quando a OpenAI alcançasse a IAG. "Achamos que será algo contínuo." As duas empresas continuariam a parceria e lançariam tecnologias cada vez mais avançadas — a um ótimo preço.

Era uma estratégia bizarra e incoerente que só fazia sentido sob uma única interpretação: a OpenAI faria o que fosse necessário, e interpretaria e reinterpretaria sua missão conforme preciso, para consolidar seu domínio.

Nos bastidores, Altman também estava preparando o terreno para consolidar seu controle. A crise no conselho deixara claro que a estrutura da OpenAI — uma organização sem fins lucrativos administrando uma com fins lucrativos — tornara sua demissão praticamente inevitável. A configuração não apenas consagrava as duas forças opostas dentro da empresa, aceleracionistas *versus* pessimistas, ambas disputando o controle do desenvolvimento da IA, como concedia ao conselho amplo poder para demiti-lo com base na interpretação deles se Altman estava ou não servindo à missão.

Tal conflito estava fadado a se repetir, caso a estrutura permanecesse

intacta.

Em 28 de maio, menos de uma semana depois de os executivos tentarem trazer Sutskever de volta, um funcionário postou novamente no canal #i-have-a-question do Slack.[5] "Não sei se devo/como perguntar isso", começou. No meio das complexidades do último acordo de acionistas da OpenAI, havia novos detalhes que pareciam permitir a dissolução da organização sem fins lucrativos. "Quão sólida é a organização sem fins lucrativos?", escreveu a pessoa. "O plano é continuar a ser governada por uma organização sem fins lucrativos?"

No dia seguinte, o *Information* relatou que esse não parecia ser o plano. Na lista de prioridades de Altman para o ano estava uma reestruturação da organização para que se parecesse mais com uma empresa típica.[6] No mês seguinte, a publicação confirmou mais detalhes. O executivo considerava alguns cenários diferentes:[7] um seria a transição da OpenAI para uma empresa tradicional com fins lucrativos; outro seria transformá-la em uma organização com fins lucrativos de utilidade pública, como a Anthropic e a xAI. Ambos os cenários manteriam a existência da organização sem fins lucrativos como uma entidade separada, mas eliminariam o controle do conselho sobre os negócios da empresa. Sob essa nova estrutura, os investidores também pressionavam Altman a adquirir ações da empresa, a fim de alinhar seus incentivos mais diretamente com os deles.

Nos meses seguintes, enquanto a OpenAI desenvolvia planos para a transição, as duas forças em jogo durante a Omnicrisis continuaram atuando. Os pessimistas intensificaram a pressão pública sobre a empresa. Em 4 de junho, o *New York Times* publicou um perfil de Daniel Kokotajlo e uma nova campanha que ele havia lançado, pedindo que as empresas de IA avançada se comprometessem com maior transparência e com a proteção de denunciantes, preservando o direito dos funcionários de alertar o público sobre os riscos que viam dentro da empresa.[8] Em uma carta aberta detalhando suas exigências, Kokotajlo foi acompanhado por doze signatários, dez deles pertencentes a diferentes épocas do clã de Segurança da OpenAI.[9] Um mês depois, o *Washington Post* noticiou que um grupo da própria OpenAI apresentara uma queixa à SEC, alegando que a empresa violara as proteções federais a denunciantes com os acordos de demissão excessivamente extensos.[10] Mais tarde no mesmo mês,[11] cinco senadores enviariam uma carta a Altman com perguntas e exigências de maior clareza sobre as alegações feitas por Leike, Kokotajlo e os demais, bem como sobre a reportagem de Piper acerca do desrespeito da OpenAI à segurança da IA e da repressão das críticas internas.

Enquanto isso, o caos na liderança e as frustrações em relação a Altman continuavam inabaláveis. Combinado com a disputa persistente entre aceleracionistas e pessimistas, isso estava levando a repetidas mudanças na estrutura hierárquica da empresa. Brockman deixou de se reportar a Murati e passou a se reportar diretamente a Altman; Aleksander Mądry foi transferido

pouco depois da carta dos senadores, deixando a liderança da equipe de Preparação para assumir uma função menor na área de pesquisa. A OpenAI também estava trazendo executivos mais experientes,[12] incluindo Sarah Friar, ex-CEO da plataforma de rede social Nextdoor, para assumir o cargo de diretora financeira, e Kevin Weil, ex-líder de produto do Facebook, Instagram e X, para ser o novo diretor de Produtos.

Em pouco tempo, a empresa perderia uma série de seus executivos mais antigos. O primeiro a sair foi John Schulman,[13] que anunciou o fato em 5 de agosto de 2024, mencionando o desejo de "aprofundar meu foco no alinhamento da IA" e a decisão de fazer isso na Anthropic. No mesmo dia, Brockman informou que tiraria um período sabático até o fim do ano,[14] explicando a decisão, em parte um desfecho das queixas dos funcionários sobre sua liderança, como uma pausa necessária após nove anos de trabalho intenso na empresa.

No mês seguinte, em 25 de setembro, a outra executiva que havia manifestado sérias preocupações sobre Altman ao conselho, Mira Murati, anunciou abruptamente que também estava saindo.[15] "Os meus seis anos e meio com a equipe da OpenAI foram um privilégio extraordinário", escreveu ela, agradecendo a Altman e Brockman e dizendo o quanto apreciava a empresa e continuaria torcendo por ela. "Estou saindo porque quero criar o tempo e o espaço para fazer minha própria exploração." Em poucas horas, outros dois líderes importantes divulgaram as próprias declarações: o diretor de pesquisa Bob McGrew e o vice-presidente Barret Zoph, que havia liderado a equipe de pós-treinamento com Schulman.[16] Os três enfatizaram, tanto aos colegas quanto ao público, que aquele era o momento certo para deixar a OpenAI, após a empresa atingir mais um marco importante: o lançamento do modelo mais recente, Strawberry, em meados de setembro, sob a nova convenção de nomenclatura, o1,[17] baseada em uma das últimas contribuições de Sutskever para a empresa.

Na verdade, o momento era péssimo. Após a Omnicrisis, a concorrência enfrentada pela OpenAI só havia acelerado. Musk expandia sua capacidade computacional a um ritmo alarmante, para construir a xAI.[18] A versão mais recente do Claude, da Anthropic, estava afastando clientes do ChatGPT.[19] Sutskever havia formado oficialmente sua nova empresa, a rival Safe Superintelligence, e acabara de anunciar um investimento inicial de um bilhão de dólares.[20]

Ao mesmo tempo, após mais de um ano de trabalho, a OpenAI ainda tinha dificuldades para alcançar o desempenho desejado a fim de justificar o lançamento do Orion.[21] A empresa começava a encarar uma perspectiva desconfortável: a fórmula comprovada de escalabilidade já não parecia ser suficiente; para desenvolver ainda mais seus sistemas de IA, seriam necessárias ideias de pesquisa fundamentalmente novas. Em geral, era muito mais fácil

falar do que fazer, especialmente depois de a OpenAI ter passado anos orientando as contratações e a própria organização das equipes para explorar pesquisas existentes, em vez de investigar territórios científicos desconhecidos.

Dois dias antes do anúncio de Murati, Altman publicou seu post mais bombástico até então, em meio à última campanha de arrecadação de fundos da OpenAI.[22] O texto, intitulado "A era da inteligência",[23] trazia promessas empolgantes sobre a prosperidade "inimaginável" que estava por vir. "Como chegamos à iminência do próximo salto em prosperidade?", escreveu. "Em dezoito palavras: a aprendizagem profunda funcionou e melhorou previsivelmente com a escala, e dedicamos cada vez mais recursos a ela."

Durante uma reunião com os funcionários,[24] Murati explicou a natureza repentina do anúncio. "Eu queria que a notícia de minha saída chegasse a todos vocês primeiro por mim, e não por meio de seus gerentes, de qualquer outra pessoa e muito menos da imprensa." Com toda a atenção voltada para a OpenAI, ela não viu como fazer isso sem que fosse uma surpresa para todos. Ao saber da decisão de Murati, McGrew decidiu também que era hora de partir: "Percebi que, na verdade, realizei a maioria das coisas importantes que queria quando vim para cá."

Assim como acontecera com a saída de Sutskever, a OpenAI procurou amenizar as três novas despedidas. "Mira, Bob e Barret tomaram essas decisões de forma independente e amigável", escreveu Altman em nota interna que depois publicou no X: "o timing da decisão de Mira fez com que fizesse sentido fazer tudo de uma vez agora, para que possamos trabalhar juntos na transição suave da para a próxima geração de liderança."[25]

No lugar de McGrew, Mark Chen, um dos líderes de pesquisa que apresentou ao lado de Murati e Zoph o evento de demonstração do 4o, assumiria o cargo de vice-presidente sênior de pesquisa, passando a liderar a divisão ao lado de Pachocki. Outro pesquisador de longa data da OpenAI, Joshua Achiam, assumiria uma nova função, a de líder de alinhamento de missões. Liam Fedus, um dos funcionários do Google que havia chegado com Zoph, lideraria o pós-treinamento. Por enquanto, contou Altman, a OpenAI não procuraria outro diretor técnico para substituir Murati.

Com todas as mudanças na liderança e na estrutura planejadas, a empresa revelava sua única constante: era e continuaria sendo o império de IA de Sam Altman.

No início de outubro de 2024, a mais recente rodada de financiamento da OpenAI foi encerrada em 6,6 bilhões de dólares, a maior rodada de capital de risco da história, avaliando a empresa em 157 bilhões. Havia uma condição: os investidores poderiam exigir seu dinheiro de volta se a empresa não se convertesse em uma entidade com fins lucrativos dentro de dois anos.

Ao longo do mesmo ano, a OpenAI continuou a perder funcionários essenciais: Luke Metz, o terceiro ex-Google que havia se juntado a Zoph e Fedus; Miles Brundage, líder de pesquisa em políticas; Lilian Weng, que

herdara o trabalho de confiança e segurança de Dave Willner e havia sido recentemente promovida a vice-presidente de pesquisa em segurança; e Alec Radford, o pesquisador original que colocou a OpenAI no caminho dos modelos GPT.

A Anthropic lançou uma campanha publicitária para o Claude com uma mensagem provocadora estampada em outdoors por São Francisco: "Aquele, mas sem todo o drama." Brockman retornou antecipadamente de seu período sabático em meio à fuga de talentos. Annie enviou a Sam uma notificação extrajudicial, informando a intenção de processá-lo. Musk, aliado do recém-reeleito presidente Donald Trump, intensificou o processo judicial contra a OpenAI, questionando a conversão planejada para uma empresa com fins lucrativos e divulgando mais e-mails antigos sobre a fundação da organização, incluindo conteúdos em que Altman falava mal de Brockman e Sutskever pelas costas ("Admitiu que perdeu muita confiança de Greg e Ilya com esse processo. Sentiu que suas mensagens eram inconsistentes e, por vezes, infantis.") e expressava aversão à transparência ("Sentiu que isso distraía a equipe.").[26] Em uma aliança surpreendente, Zuckerberg, historicamente rival de Musk, apoiou-o. Em uma carta ao procurador-geral da Califórnia, a Meta solicitou igualmente que a conversão fosse bloqueada. "A conduta da OpenAI pode ter implicações sísmicas para o Vale do Silício", declarou a empresa.[27] A conversão poderia criar um precedente perigoso, incentivando outras startups a se registrarem como organizações sem fins lucrativos, garantindo a si mesmas e a seus investidores vantagens fiscais do governo até o momento em que se tornassem lucrativas.

Em dezembro, em meio às notícias típicas das festas de fim de ano, a OpenAI anunciou oficialmente os planos para a nova estrutura.[28] A organização passaria a operar como uma empresa com fins lucrativos de utilidade pública, enquanto a entidade sem fins lucrativos continuaria existindo separadamente, mantendo participação acionária na nova empresa. Segundo o anúncio, essa configuração seria a melhor forma de dotar tanto a parte com fins lucrativos quanto a sem fins lucrativos dos recursos necessários para cumprir os objetivos enquanto seguia a missão da OpenAI. "Mais uma vez, precisamos levantar mais capital do que imaginávamos", dizia o comunicado. "O mundo está avançando para construir uma nova infraestrutura de energia, uso do solo, chips, data centers, dados, modelos e sistemas de IA para a economia do século XXI. Buscamos evoluir para dar o próximo passo em nossa missão, ajudando a construir a economia da IAG e garantindo que ela beneficie a humanidade."

No início do novo ano, Altman voltou a fazer declarações grandiosas. "Estamos confiantes de que sabemos como construir a IAG, tal como tradicionalmente a entendemos", escreveu ele em uma nova postagem no blog em 6 de janeiro de 2025. "Estamos começando a direcionar nosso foco para além disso, para a superinteligência, no verdadeiro sentido da palavra."[29]

EPÍLOGO

# CASTELO DE CARTAS

Em 2021, deparei-me com uma história diferente de todas as que já tinha relatado: a história de uma comunidade indígena na Nova Zelândia que estava usando IA para revitalizar o *te reo Māori*, a língua tradicional do povo maori.[1]

Assim como muitos grupos indígenas em todo o mundo, gerações dos maoris enfrentaram um tratamento horrível sob o domínio colonial. Em 1867, sob a Lei de Escolas Nativas, que tornou o inglês a única língua permitida nas salas de aula da Nova Zelândia, crianças maoris eram humilhadas e até espancadas por falarem a própria língua. Depois, com a rápida urbanização que varreu o país no início do século XX, as comunidades maoris se desintegraram e se dispersaram, enfraquecendo seus centros de preservação cultural e linguística. O número de fluentes em *te reo* caiu de 90% para 12% da população maori. Quando a Nova Zelândia, ou Aotearoa, como os maoris originalmente chamavam sua terra, reverteu essas políticas 120 anos depois, restavam poucos professores do idioma para ressuscitar uma língua em extinção. Como tantas outras antes dela, o *te reo* quase desapareceu da face da Terra.

É difícil transmitir plenamente a tragédia que é perder uma língua. Pelas mesmas razões pelas quais os pesquisadores de IA se voltaram no início para a linguagem ao desenvolver as tecnologias, a extinção de um idioma representa muito mais do que a perda de um meio de comunicação. Cada um carrega uma história, uma cultura e um corpo de conhecimento; é o produto coletivo de milhões de pessoas ao longo do tempo, buscando sons e formas escritas para capturar as observações mais sutis sobre o universo, sobre a vida, sobre a experiência humana; para compartilhar uns com os outros a beleza deslumbrante e o fracasso doloroso; para ensinar uma criança, aprender com um ancião; para expressar amor.

Perder uma língua é uma tragédia global, e profundamente pessoal. Ser separado da própria herança e forçado a preservar a de outra pessoa, ou correr o risco de ser espancado, expõe, da forma mais crua possível, uma hierarquia explícita a respeito de quais histórias, culturas e conhecimentos merecem ser transmitidos e quais são considerados insignificantes e podem ser apagados.

Os grandes modelos linguísticos aceleram a perda de línguas. Mesmo para

os de gerações anteriores, como o GPT-2, apenas poucos idiomas no mundo têm falantes em número suficiente e documentação on-line em escala adequada para suprir a necessidade de dados desses sistemas.[2] Entre os mais de 7 mil ainda existentes, quase metade está em perigo de extinção, segundo a Unesco;[3] cerca de um terço tem alguma presença on-line; menos de 2% são encontrados no Google Tradutor; e, de acordo com os testes da própria OpenAI, apenas quinze, ou 0,2%, são atendidos pelo GPT-4 com precisão superior a 80%. À medida que esses modelos se tornam infraestrutura digital, a acessibilidade da internet para diferentes comunidades linguísticas, e, com ela, o acesso às oportunidades econômicas que a internet oferece, continuará diminuindo, incentivando cada vez mais essas comunidades a priorizarem o aprendizado e o uso de um idioma dominante, como o inglês, em detrimento do próprio.

Foi diante dessa ameaça existencial iminente, uma concepção fundamentalmente diferente do que costuma ser chamado de existencial,[4] que um casal indígena, Peter-Lucas Jones e Keoni Mahelona, recorreu pela primeira vez à IA como possível ferramenta para ajudar uma nova geração de praticantes do idioma a devolver ao *te reo* seu vigor. Jones, que é maori, e Mahelona, que é nativo do Havaí, são parceiros na vida e no trabalho. Os dois se conheceram e se apaixonaram, conta Mahelona, depois de ele ter uma visão em um sonho: caso se mudasse para a Nova Zelândia, conheceria um rapaz maori com quem compartilharia a vida.

Em 2012, os dois se mudaram de Wellington para Kaitāia, cidade natal de Jones, no extremo norte de Aotearoa. Ele tornou-se CEO da Te Hiku Media, uma emissora pública que transmite em *te reo* e integra uma rede mais ampla de meios de comunicação e outras organizações dedicadas à revitalização da língua. Na nova função, Jones identificou uma oportunidade. Ao longo de mais de vinte anos de transmissão, a Te Hiku havia acumulado um valioso acervo de arquivos de áudio de pessoas falando *te reo*, incluindo uma gravação da própria avó, Raiha Moeroa, nascida no fim do século XIX, cujo sotaque ainda não havia sido distorcido pelas influências do inglês colonial. Jones também tinha a ambição de registrar muitas outras entrevistas com anciãos maoris, a fim de documentar suas histórias orais e o *te reo* nativo antes que falecessem. Essas gravações, acreditava ele, poderiam se tornar um recurso precioso para o aprendizado da língua, um portal para o passado capaz de permitir que novas gerações ouvissem os sons originais do idioma e se conectassem com a sabedoria dos antepassados.

O desafio era transcrever o áudio para ajudar os alunos a acompanhar, dada a escassez de pessoas fluentes em *te reo*. Assim, em 2016, quando a OpenAI estava começando, Jones recorreu a Mahelona, que, naquele momento, reformulava o site da Te Hiku, para encontrar uma solução. Polímata, estudou engenharia mecânica na Olin College, gestão de negócios em seu primeiro mestrado e física e nanotecnologia computacional no segundo, então Mahelona, como bolsista Fulbright (um dos programas de

intercâmbio acadêmico mais prestigiados do mundo, patrocinado pelos Estados Unidos) na Nova Zelândia, rapidamente teve a ideia de usar IA: com um modelo de reconhecimento de voz em *te reo* cuidadosamente treinado, a Te Hiku seria capaz de transcrever seu vasto repositório de áudio contando apenas com pessoas fluentes no idioma.

É aqui que a história da Te Hiku diverge completamente do modelo de desenvolvimento de IA da OpenAI e do Vale do Silício.[5] Conscientes dos efeitos devastadores da expropriação colonial, Jones e Mahelona estavam determinados a avançar com o projeto somente se pudessem garantir três princípios fundamentais: consentimento, reciprocidade e soberania maori em todas as etapas do desenvolvimento. Isso significava que, antes mesmo de iniciar o projeto, buscariam a permissão da comunidade maori e de seus anciãos, perguntando se aquele empreendimento era realmente desejado; que, para coletar dados, aceitariam contribuições apenas de pessoas que compreendessem plenamente como eles seriam usados e estivessem dispostas a participar; que, para maximizar os benefícios do modelo, ouviriam a comunidade para compreender quais tipos de recursos de aprendizagem de idiomas seriam mais úteis; e que, uma vez obtidos esses recursos, comprariam as próprias GPUs e os próprios servidores Nvidia para treinar os modelos localmente, sem depender da infraestrutura em nuvem de grandes empresas de tecnologia.

Mais importante ainda, a Te Hiku estabeleceria um processo no qual os dados coletados se tornariam um recurso voltado para benefícios futuros, sem jamais poder ser apropriado por projetos que, sem o consentimento da comunidade, pudessem explorá-los, prejudicá-los ou violar seus direitos. Com base no princípio maori de *kaitiakitanga*, ou tutela, esses dados permaneceriam sob a administração da Te Hiku, em vez de serem publicados on-line. A organização concederia licenças apenas a entidades que respeitassem os valores maoris e que pretendessem utilizar o material em projetos aprovados e considerados úteis pela comunidade.

"Os dados são a última fronteira da colonização",[6] falou Mahelona. Os antigos impérios tomaram as terras das comunidades indígenas e depois as obrigaram a comprá-las de volta, sob termos e serviços restritivos, caso quisessem recuperar a própria propriedade. "A IA é apenas mais uma apropriação de terras. As grandes empresas de tecnologia gostam de recolher dados praticamente de graça, para construir o que quiserem, seja qual for o objetivo final, e depois devolvê-los a você como um serviço pago."

Do início ao fim, Jones e Mahelona conduziram o projeto sem fazer concessões. Em determinado momento, lançaram uma campanha educativa para ensinar mais maoris sobre IA e criaram uma competição comunitária para incentivar doações de dados e anotações. Em apenas dez dias, a Te Hiku reuniu 310 horas de áudio transcrito de alta qualidade, provenientes de cerca de duzentas mil gravações feitas por aproximadamente 25 mil pessoas. O nível de envolvimento foi inédito entre muitos pesquisadores de IA, uma prova da

confiança e do entusiasmo que a abordagem da Te Hiku despertou na comunidade. As pessoas ficaram mais do que dispostas a doar os dados assim que compreenderam o projeto e concordaram com ele, certas de que a Te Hiku continuaria a administrar esse material de forma responsável.

Esse conjunto de dados não foi nada diante das 680 mil horas de áudio que a OpenAI extraiu da internet para treinar sua ferramenta de reconhecimento de voz,[7] Whisper. No entanto, a experiência da Te Hiku oferece outra lição importante: as 310 horas coletadas foram suficientes para desenvolver o primeiro modelo de reconhecimento de voz em *te reo*, alcançando 86% de precisão.

Enquanto a OpenAI busca criar modelos únicos, massivos e capazes de realizar inúmeras tarefas, um objetivo que exige o máximo de dados possível, a Te Hiku concentrou-se simplesmente em desenvolver um modelo pequeno e especializado, excelente para uma função específica. Além disso, beneficiou-se da comunidade transfronteiriça e de código aberto: tomou como ponto de partida o DeepSpeech,[8] um modelo gratuito da Fundação Mozilla, ele próprio fruto de uma visão distinta de desenvolvimento em IA. Assim como a Te Hiku, a Mozilla treinou o modelo apenas com dados doados mediante consentimento total, utilizando uma arquitetura de rede neural criada pelo laboratório de pesquisa da empresa chinesa Baidu, sediada em São Francisco. No total, a Te Hiku utilizou apenas duas GPUs.

Escrevi sobre o trabalho da Te Hiku antes de o ChatGPT se apoderar do paradigma dominante de desenvolvimento de IA, jogando pela janela o consentimento, a reciprocidade e a soberania. Contudo, nos anos que se seguiram, passei a ver a abordagem radical da Te Hiku como ainda mais relevante e vital. As críticas que apresento neste livro à visão mais ampla da OpenAI e do Vale do Silício não têm, de forma alguma, a intenção de descartar a ferramenta por completo. O que rejeito é a noção perigosa de que os amplos benefícios da IA só podem ser derivados — na verdade, *só* surgirão — de uma visão da tecnologia que exige a cessão completa de nossa privacidade, de nossa autonomia e de nosso valor, incluindo o valor de nosso trabalho e de nossa arte, em prol de um projeto de centralização imperialista.

A Te Hiku nos mostra outro caminho. Ela imagina como a IA e seu desenvolvimento poderiam ser exatamente o oposto do que hoje se tornou dominante. Os modelos podem ser pequenos e específicos para determinadas tarefas, com dados de treino contidos e reconhecíveis, eliminando os incentivos para práticas laborais exploratórias e psicologicamente prejudiciais, bem como o extrativismo intensivo em recursos acompanhando a produção e o funcionamento de supercomputadores. A criação de inteligência artificial pode ser impulsionada pela comunidade, baseada em consentimento e respeitosa com relação ao contexto e à história locais; suas aplicações podem elevar e fortalecer comunidades marginalizadas; e sua governança pode ser inclusiva e democrática.

A Te Hiku não é a única organização a buscar novos caminhos para o desenvolvimento da IA. Ao longo de minha pesquisa para este livro, encontrei repetidamente organizações e movimentos ao redor do mundo que surgiram para resistir aos impérios da IA, afirmar seus direitos à autodeterminação e imaginar novas trilhas possíveis.

Depois de Timnit Gebru ter sido demitida do Google,[9] fundou uma organização sem fins lucrativos em dezembro de 2021 para continuar a pesquisa.[10] Ela a chamou de DAIR, o Instituto de Pesquisa em IA Distribuída, "distribuída" para desafiar a centralização. "Essa foi a primeira palavra que me veio à mente", diz Gebru. Ela imaginou formar uma equipe de pesquisadores ao redor do mundo que permanecessem inseridos nas próprias comunidades, trazendo para o trabalho do instituto as ricas experiências e perspectivas de seus contextos locais, e usando o que foi aprendido para beneficiar essas mesmas comunidades. "A tecnologia está impactando o mundo inteiro fora do Vale do Silício, mas o mundo inteiro não está tendo a oportunidade de impactar a tecnologia", afirma ela.

Alex Hanna, sociólogo e um dos coautores do artigo "Papagaios Aleatórios", do Google, foi o primeiro a se juntar a Gebru como diretor de pesquisa da DAIR. Sua primeira tarefa foi elaborar uma filosofia que expressasse com mais clareza o ethos da organização. Para isso, Gebru e Hanna contrataram a terceira integrante da DAIR, Milagros Miceli, também socióloga e cientista da computação, que vinha pesquisando as práticas laborais exploratórias da indústria de IA. Juntos redigiram a seguinte filosofia: "Nossa investigação tem como objetivo beneficiar comunidades que normalmente não são atendidas pela IA e criar caminhos para recusar, questionar e remodelar os sistemas de IA em conjunto."[11]

Eles criaram sete pilares para orientar a implementação dessa filosofia, o que inclui centralizar e construir relações significativas com as comunidades afetadas, mas ainda não representadas na pesquisa em IA, tratando-as como parceiras reais na produção de conhecimento; compensar de forma justa qualquer trabalho envolvido na criação de pesquisas e tecnologias; questionar os sistemas que sustentam o desenvolvimento da IA e que continuam a marginalizar aqueles que sempre foram historicamente marginalizados; e trabalhar com essas comunidades para imaginar alternativas que possam, pouco a pouco, remodelar o mundo em direção a um lugar onde elas realmente queiram viver.

A partir daí, Miceli iniciou um novo projeto de pesquisa para colocar essa filosofia em prática. Criou a Data Workers' Inquiry (ou Investigação dos Trabalhadores de Dados) e convidou trabalhadores de dados de diversas partes do mundo a formular as próprias perguntas de pesquisa sobre a indústria de anotação de dados e sobre como torná-la melhor. Independentemente do país em que viviam, ela os remunerou com o salário padrão de um pesquisador na Alemanha,[12] onde está sediada, para refletir o real valor do trabalho que realizavam: 25 euros por hora.

"Há sempre essa lógica falsa em torno do trabalho com dados: qual é o mínimo que podemos pagar a essas pessoas? Isso vem de uma lógica colonialista: você escolhe um lugar que permita fazer o máximo com o orçamento mais baixo e onde seja possível, de fato, roubar pessoas e extrair recursos a custo mínimo", diz Miceli. "A questão é: por que essas empresas pagam dois dólares por hora se esse trabalho lhes rende bilhões ou trilhões em receita? Por que não olhamos para quanto essas empresas *podem* pagar, em vez de para o mínimo que esses trabalhadores podem aceitar?"

Entre os quinze trabalhadores que participaram da primeira rodada da investigação estavam Oskarina Veronica Fuentes Anaya, da Venezuela, e Mophat Okinyi, do Quênia. Para o projeto,[13] Fuentes fez parceria com um artista de animação para criar um vídeo sobre suas experiências e colaborou com outros trabalhadores de dados para destacar os desafios que todos compartilham: a escassez de tarefas nas plataformas, a imprevisibilidade e a falta de controle sobre os horários de trabalho, além da remuneração miserável. Hoje, ela trabalha simultaneamente em cinco plataformas de anotação de dados para conseguir ganhar um pouco mais que o salário mínimo na Colômbia, cerca de 335 dólares por mês. Cada tarefa paga, em média, entre um e cinco centavos; ainda assim, ela se obriga a acordar quando as tarefas chegam no meio da noite. "Somos fantasmas para a sociedade e, ouso dizer, mão de obra barata e descartável para as empresas que servimos há anos sem garantias ou proteção", escreveu em seu projeto. Desde a Data Workers' Inquiry, ela segue compartilhando experiências em palestras on-line e webinars, na esperança de pressionar empresas e os formuladores de políticas a assegurar um tratamento melhor aos trabalhadores.

A um continente de distância, Okinyi também estava se organizando.[14] Em maio de 2023, pouco mais de um ano após o término abrupto do contrato da OpenAI com a Sama, ele tornou-se organizador da União Africana de Moderadores de Conteúdo, sediada no Quênia, que luta por melhores salários e condições para os trabalhadores africanos que realizam o pior trabalho da internet. Seis meses depois, após tornar pública sua experiência na OpenAI por meio de meu artigo no *Wall Street Journal*, ele fundou a própria organização sem fins lucrativos, a Techworker Community Africa (TCA),[15] ao lado de um de seus ex-colegas da Sama, Richard Mathenge.

Em agosto de 2024, quando nos encontramos, Okinyi imaginava transformar a TCA em um recurso tanto para a comunidade africana de trabalhadores de dados de IA quanto para grupos internacionais e formuladores de políticas que buscavam apoiá-los. Havia organizado conferências on-line e palestras em escolas a fim de ensinar trabalhadores e estudantes, especialmente mulheres, sobre direitos trabalhistas, privacidade de dados e funcionamento interno da indústria de IA. E, como buscava financiamento para abrir um centro de formação que qualificasse mais pessoas, reuniu-se com representantes dos Estados Unidos que visitaram Nairóbi para entender melhor a experiência dos trabalhadores que prestam serviços a

empresas de tecnologia norte-americanas. Também estava atendendo diversos pedidos de organizações mundiais, entre elas a Equidem, dedicada à defesa de direitos humanos e trabalhistas no Sul Global, e o projeto Fairwork, do Oxford Internet Institute.

Para a Data Workers' Inquiry, ele entrevistou trabalhadores da Remotasks, no Quênia, cujo acesso à plataforma foi bloqueado pela Scale (companhia de tecnologia dos Estados Unidos), fazendo desaparecer os últimos rendimentos que eles não haviam conseguido resgatar. Ele utilizou parte das doações arrecadadas pela TCA para apoiá-los durante esse pesadelo financeiro. "À medida que a poeira baixava, uma coisa permanecia: o custo humano do poder descontrolado e da ganância desenfreada",[16] escreveu ele. "As vozes desses trabalhadores ecoam a esperança de um futuro mais brilhante e mais equitativo… é um apelo à ação para garantir que trabalhadores em todo o mundo sejam tratados com a dignidade e o respeito que merecem."

Segundo Okinyi, a dignidade e o respeito que recebeu pela atuação fortaleceram-no com uma nova esperança e melhoraram profundamente sua saúde mental. Pouco antes de nossa conversa, ele recebeu a notícia de que integraria a lista anual da *Time* das cem pessoas mais influentes em IA. "Sinto que meu trabalho está sendo valorizado",[17] comenta. Isso não significa que o caminho tenha sido fácil. Em março de 2024, ele pediu demissão do emprego em tempo integral que assumira após sair da Sama. Segundo relata, a liderança da empresa não via com bons olhos seu trabalho de organização: "Eles achavam que eu influenciaria os funcionários a se tornarem ativistas." A mesma empresa transferiu alguns dos projetos para Gana à medida que o sindicato e a TCA se tornavam mais atuantes. Ele soube, depois, que autoridades quenianas reclamaram que a mobilização dos trabalhadores estaria afastando investimentos e deixando mais pessoas desempregadas no país.

A abrangência global característica da indústria tornou Okinyi ainda mais determinado a chamar a atenção internacional para os trabalhadores de dados africanos. Mesmo que o governo queniano fosse favorável, a legislação do país, por si só, pouco poderia fazer para restringir o comportamento das empresas de IA. A maioria delas vem dos Estados Unidos, especificamente de São Francisco, diz ele. Por isso, é necessário um esforço internacional coordenado para responsabilizá-las.

No Uruguai, Daniel Pena chega às mesmas conclusões.[18] A cadeia de abastecimento da indústria de IA é complexa e extensa. "Eles retiram energia daqui, os dados vão para lá; extraem minerais de algum lugar, trazem trabalhadores de outro lugar", explica. Diante desses impactos crescentes e das gigantes tecnológicas por trás deles, cada comunidade que trava sua batalha local pode sentir-se isolada e impotente, especialmente quando é prejudicada pelos próprios governos, que "precisam das empresas para manter a aparência de uma economia estável". Pouco depois de conhecê-lo, descobriu que o próprio governo havia ignorado sua petição, assinada por mais de quatrocentas pessoas, pedindo um estudo mais aprofundado sobre os impactos sociais e

ambientais do data center do Google no país. Em vez disso, o Ministério do Meio Ambiente aprovou discretamente o projeto, revelando a decisão apenas depois de expirado o prazo de trinta dias para contestação pública. Ainda assim, Pena não pretende desistir. Ele tem conversado com a Mosacat, no Chile, e buscado contato com o maior número de outras comunidades que também resistem à exploração e ao extrativismo da indústria tecnológica. Ao conectar esses movimentos que ultrapassam fronteiras, compartilhando informações e estratégias de resistência, ele vê um caminho para construir um poder coletivo capaz de pressionar e transformar a indústria em algo melhor. "Precisamos lutar em nível global", afirma.

Se a missão da OpenAI é uma fórmula para construir um império, qual seria então a fórmula para dissolvê-lo? Enquanto escrevo este livro, é impossível prever em detalhes como a empresa, ou a indústria de IA em rápida evolução, continuará a se desenvolver. Talvez um dos muitos concorrentes da OpenAI acabe substituindo sua posição de liderança; muito provavelmente, os impérios da inteligência artificial ajustarão as táticas, reinventando a forma como desenvolvem modelos, exploram mão de obra e expandem a infraestrutura computacional. Contudo, independentemente de como os próximos dois ou dez anos se desenrolem, há medidas que precisamos tomar e que não devem mudar.

Numa palestra de 2019 na NeurIPS,[19] durante o workshop Queer in AI, Ria Kalluri, pesquisadora de IA em Stanford, propôs uma alternativa incisiva à ideia de garantir que a IA faça o "bem". "Bondade", "benefício para a humanidade", esses termos sempre estarão nos olhos de quem os interpreta. Em vez disso, argumentou ela, devemos perguntar como a inteligência artificial altera as relações de poder: ela consolida esse poder ou o redistribui? No contexto deste livro, a questão torna-se ainda mais clara: a IA continua a fortalecer o império ou começa a nos conduzir de volta à democracia?

Dirigindo-se a um público técnico, Kalluri concentrou a palestra na pesquisa fundamental em IA, em como os cientistas poderiam usar essa pergunta para avaliar que tipos de sistemas construir e quais direções seguir para avançar no campo. Contudo, sua formulação é igualmente crucial para todos os outros aspectos da IA. Como deveríamos desenvolver aplicações de IA, como deveríamos utilizá-las e, em última instância, conforme perguntei no início deste livro, como poderíamos governar essa tecnologia de modo a devolver o poder às pessoas?

O trabalho de Te Hiku, da DAIR, de Okinyi, Fuentes e Pena exemplifica o tipo de iniciativa que pode, e precisa, ser feita para redistribuir o poder. Contudo, a questão da governança diz respeito a como criar as condições necessárias para que mais trabalhos iguais a esses possam proliferar e florescer.

Na palestra, Kalluri propôs a ideia de diferentes eixos de poder. Este livro aborda três deles: conhecimento, recursos e influência. Atualmente, a OpenAI e seus impérios concorrentes controlam cada um desses eixos. Por meio da

centralização de talentos, da erosão da ciência aberta e do isolamento de modelos do escrutínio público, controlam a produção de conhecimento; ao acumular financiamento, dados, mão de obra, computação, energia e terra, controlam recursos e reduzem o acesso de outras pessoas a eles; e, ao criar e reforçar ideologias e produzir demonstrações populares que capturam a imaginação mundial, exercem influência de longo alcance. Cada um desses fatores alimenta o outro. Controlar a produção de conhecimento fortalece a influência; uma influência crescente atrai mais recursos; e o acúmulo de recursos garante a capacidade de continuar produzindo conhecimento.

A fórmula para dissolver o império exige, portanto, a redistribuição do poder ao longo dos eixos. As sugestões e recomendações que apresento aqui são alguns exemplos, mas de modo algum abrangentes. Em primeiro lugar, para redistribuir o conhecimento, precisamos de mais financiamento para apoiar a produção fora do império. Isso significa sustentar pesquisadores capazes de realizar avaliações independentes dos modelos corporativos, para que não dependamos das empresas para compreender suas capacidades. Significa apoiar organizações como a DAIR, que podem explorar direções de pesquisa completamente novas, incluindo formas de IA além dos grandes modelos de linguagem, mais eficientes em dados e energia. Significa também apoiar organizações como a Te Hiku, que desenvolvem aplicações de IA específicas para determinadas tarefas e orientadas para a comunidade, fortalecendo grupos historicamente marginalizados. A produção independente de conhecimento inclui, ainda, o trabalho de jornalistas e grupos da sociedade civil que conseguem se integrar às comunidades e estar presentes no terreno, ajudando-nos a compreender, em vez de apenas especular, as realidades complexas do impacto dessas tecnologias.

A redistribuição do conhecimento também exigirá políticas que obriguem as empresas a divulgar detalhes essenciais sobre os dados de treinamento e as especificações técnicas de modelos e supercomputadores. Só assim avaliadores independentes poderão realizar o trabalho. A pesquisadora Deborah Raji,[20] da Universidade da Califórnia, Berkeley, que continuou colaborando com formuladores de políticas após os fóruns de Schumer, afirma que esse também é o mínimo necessário para garantir a segurança real dos sistemas corporativos. Ou seja, não os danos teóricos de uma IA pessimista, mas os danos concretos já existentes: discriminação, desinformação e automação do trabalho, que consumidores e comunidades enfrentam quando modelos amplamente implantados não passam por testes adequados. "Temos o CFPB (Consumer Financial Protection Bureau), que monitora produtos financeiros. Temos a FDA, que monitora dispositivos médicos. Por alguma razão, porém, quando se trata de produtos de IA, simplesmente não há fiscalização", diz Raji. Os modelos deveriam apresentar mais transparência do que produtos no geral. "Estes são sistemas definidos por dados. Não são determinísticos. Portanto, precisamos saber mais sobre esses sistemas para entender o que estão fazendo."

Essa transparência é ainda mais crucial para avaliar o impacto ambiental da inteligência artificial. Nesse aspecto, outros produtos já passam por avaliações regulatórias às quais os produtos de IA não são submetidos. "Se você usa um carro, se compra um eletrodoméstico, ele vem com uma classificação Energy Star [certificação internacional de eficiência energética]", [21] declara Sasha Luccioni, da Hugging Face. "Mas a IA está tão integrada à nossa sociedade, tão amplamente utilizada em produtos, e ainda assim não temos informação sobre a sustentabilidade desses sistemas."

Com essa transparência, também começaríamos a redistribuir o poder ao longo de nosso segundo eixo: os recursos. Ao ocultar os componentes dos modelos sob a justificativa de propriedade intelectual, os impérios da IA têm conseguido, até este momento, apropriar-se da propriedade intelectual de outras pessoas sem oferecer crédito, pedir consentimento ou dar compensação. A visibilidade sobre os dados de treinamento das empresas tornaria esse comportamento extrativista e exploratório muito mais difícil. O mesmo ocorreria com a transparência sobre as cadeias de abastecimento, incluindo onde contratam a mão de obra e onde negociam novos arrendamentos de terra para construir mais usinas elétricas e centros de dados, algo que frequentemente acontece por meio de entidades de fachada.

A redistribuição de recursos também exige proteções trabalhistas mais fortes em todos os níveis, não apenas para os trabalhadores de dados contratados diretamente pela indústria, mas para todos aqueles cujas produções correm o risco de ser cooptadas como dados de treino ou cujos empregos podem ser automatizados. As greves de Hollywood, que garantiram com sucesso proteções para escritores e atores contra determinados usos da IA, ilustraram o papel fundamental que os sindicatos desempenharão na resistência à desvalorização do trabalho humano, à redução dos salários e à concentração do dinheiro nas mãos das empresas de IA.

Por fim, para redistribuir o poder ao longo do nosso terceiro eixo, a influência, precisamos de educação abrangente. O antídoto para o misticismo e a ilusão do hype da IA é ensinar às pessoas como ela funciona, quais são os pontos fortes e fracos, quais sistemas moldam o desenvolvimento e quais visões de mundo e limitações têm as pessoas e empresas que constroem essas tecnologias. Como afirmou Joseph Weizenbaum, professor do MIT e inventor do chatbot ELIZA na década de 1960: "Uma vez que determinado programa é desmascarado, uma vez que o funcionamento interno é explicado em linguagem suficientemente fácil para induzir a compreensão, a magia se desfaz."[22] Espero que este livro seja apenas uma contribuição para ampliar essa compreensão. Ele se apoia no trabalho de muitos estudiosos, jornalistas, ativistas e educadores que vieram antes de mim e se dedicaram à educação do público. Que sirva de terreno para que muitos outros, depois de mim, se levantem e construam.

## POSFÁCIO

As pessoas com frequência me perguntam como a situação mudou desde que escrevi a primeira edição deste livro. Em muitos aspectos, a validade da minha analogia com impérios históricos se fortaleceu a uma velocidade que foi surpreendente até para mim. Desde sua posse, Donald Trump deu todo o apoio do governo dos Estados Unidos ao respaldar o setor norte-americano de inteligência artificial, impulsionando a OpenAI e seus concorrentes a um ritmo vertiginoso, mesmo para seus próprios padrões. No início de 2025, Altman anunciou oficialmente o projeto Stargate, atualizado, de sua concepção inicial como um único supercomputador, para um plano de investimento de 500 bilhões de dólares ao longo de quatro anos para o desenvolvimento de uma infraestrutura computacional muito mais ampla. Na época, esse número já parecia fantasioso. Em comparação, o programa espacial Apollo, da NASA, que levou o primeiro homem à Lua, custou cerca de 300 bilhões (ajustados para o dólar de 2025), durante treze anos. No entanto, no outono de 2025, Altman já implementava uma meta drasticamente diferente, afirmando que pretende construir data centers com 250 gigawatts de nova capacidade — exigindo algo como o equivalente energético de quase cinquenta cidades da dimensão de Nova York — até 2033. Ele estima que o projeto custará dez trilhões de dólares.

Como descrever algo de tamanha magnitude? Uma pessoa comum jamais se deparou com dez trilhões de qualquer coisa. É impossível não ver a utilização descuidada de números incompreensíveis como parte da estratégia. A esta altura, o que estamos testemunhando é uma reprise contemporânea da Companhia Britânica das Índias Orientais, cuja expansão arriscada contou com o apoio da Coroa britânica. Se continuarmos neste caminho, as consequências não serão somente a exploração laboral, os danos ambientais e as crises na saúde pública; se retornarmos à era dos impérios, a democracia não sobreviverá.

Ainda assim, ao mesmo tempo, nunca me senti tão esperançosa quanto à nossa capacidade de construir um futuro diferente.

Para minha profunda gratidão, *O império da IA* alcançou leitores em dezenas de países. Quando me encontrei com milhares deles ao redor do mundo, a energia e a convicção que vi em todos os lugares para resistir aos impérios e exigir algo melhor têm sido extraordinárias. Não há uma semana

sequer que eu não ouça falar dos novos movimentos que brotam desse terreno tão fértil. Comunidades protestam contra data centers extrativistas; artistas e escritores processam empresas de IA por roubarem sua propriedade intelectual; trabalhadores entram em greve contra a exploração laboral viabilizada pela inteligência artificial; cidadãos pressionam governos em todos os níveis para que leis e regulamentações sejam estabelecidas a respeito do uso da IA; pesquisadores e empreendedores estão imaginando concepções fundamentalmente novas dessa tecnologia que sirvam para fortalecer comunidades e proteger o planeta, e investidores estão redirecionando capital para financiar essas ideias.

Agora carrego comigo um exemplar de *Hope in the dark*, de Rebecca Solnit, um belo lembrete do poder radical e transformador da esperança. "Esperança... não é a crença de que tudo estava, está ou estará bem. As evidências de enorme sofrimento e enorme destruição estão por toda parte", escreve ela. "A esperança se situa na premissa de que não sabemos o que vai acontecer e de que, na imensidão da incerteza, há espaço para a ação. Quando alguém reconhece a incerteza, reconhece que pode ser capaz de influenciar os resultados — sozinho ou junto com algumas dezenas ou vários milhões de outros."

Sendo assim, neste grande momento de incerteza, desejo que você reconheça o papel ativo que tem a desempenhar na condução do desenvolvimento tecnológico, na resistência aos impérios e no fortalecimento do nosso futuro coletivo. Impérios são criados para parecerem inevitáveis. No entanto, a história nos mostrou: quando o povo se ergue, os impérios sempre caem.

— Karen Hao, novembro de 2025

# AGRADECIMENTOS

Um tema central deste livro é confiança. Confiança na aprendizagem profunda. Confiança na IAG. Autoconfiança. De que maneira a confiança mobiliza e impulsiona. Em quem acreditar e em quem não acreditar. A confiança é algo poderoso e intoxicante. E, na minha própria carreira, foi a confiança de tantas pessoas em mim e em meu trabalho que se tornou o meu maior facilitador.

Em primeiro lugar, agradeço às pessoas que acreditaram neste projeto de livro. Sou especialmente grata a todas as minhas fontes. Muitas falaram comigo apesar dos riscos legais ou de outras naturezas, porque acreditavam na verdade, na transparência e na responsabilidade. Muitas também foram extremamente generosas com seu tempo, convidando-me para suas casas, mostrando-me suas comunidades ou participando de mais de dez horas de entrevistas ao longo de várias sessões. Nunca subestimarei o ato de fé que alguém realiza ao abrir o coração e a mente para um jornalista. É uma honra absoluta contar suas histórias. Obrigada. Sem vocês, este livro simplesmente não existiria.

Minha sincera gratidão a David Doerrer, meu maravilhoso agente, que foi o primeiro a dedicar tempo e apoio para me ajudar a desenvolver as ideias deste livro, antes mesmo de eu ter algo digno de lhe mostrar. Foi graças à sua paciência, às suas perguntas perspicazes e, sobretudo, à sua capacidade de me dizer com gentileza e firmeza quando algo não estava funcionando que comecei a perceber como minha coleção dispersa de pensamentos poderia formar a base de uma publicação.

A Scott Moyers, da Penguin Press, o sonho de qualquer escritor em termos de editor, que compreendeu imediatamente minha visão e cujo apoio incondicional desde então me permitiu persegui-la de forma mais ambiciosa. Ele não apenas foi uma bússola moral e um incentivador ao longo de todo o meu processo de reportagem e escrita, oferecendo feedback perspicaz e esclarecedor, como também, juntamente com Ann Godoff, garantiu os recursos financeiros e jurídicos da Penguin Press para apoiar a criação desta obra. Pesquisar e escrever um livro como este é caro, delicado e demorado. Entre muitas outras coisas, é preciso contratar investigadores e verificadores de fatos; pagar voos, alojamento, colaboradores locais, tradutores e motoristas

para passar tempo em campo, integrado nas comunidades. Scott e Ann tornaram possível que eu fizesse tudo isso e me dedicasse integralmente a este livro.

A Mia Council, da Penguin Press, cujas edições afiadas e compassivas me deram o incentivo e a segurança de que eu precisava para assumir riscos e levar minha escrita a um novo patamar. Sua coordenação magistral nos bastidores fez com que todo o processo de edição e produção parecesse harmonioso. Tenho certeza de que nem cheguei a ver metade do caos logístico que Mia administrou com tanta habilidade. Agradeço também ao restante das equipes extraordinárias da Penguin Press e da Penguin Random House nos Estados Unidos e no Reino Unido, incluindo Yuki Hirose, Gail Brussel, Juli Kiyan, Danielle Plafsky, Laura Stickney, Kim Walker, Rosie Brown, Lotte Hall, Karen Dziekonski e muitos outros com quem não interagi diretamente, mas que foram essenciais para o processo.

Ao meu incrível time de verificação de fatos: Lindsay Muscato, Matt Mahoney, Rima Parikh e Muriel Alarcón. Os quatro revisaram meticulosamente o rascunho, conferindo o labirinto de detalhes com documentos e fontes e testando cada escolha de palavra. Matt também apoiou a pesquisa inicial do livro, enquanto Lindsay atendeu inúmeras chamadas minhas, tornando-se a ouvinte mais paciente possível. Rima, por sua vez, transformou suas notas de verificação de fatos em um verdadeiro stand-up. Todos são salvadores.

Muriel também foi minha parceira extraordinária de reportagem no Chile e no Uruguai. Ela é uma mulher maravilhosa: conduziu pesquisas, coordenou entrevistas, buscou fontes e atuou como tradutora e motorista durante duas semanas de apuração ininterrupta, tudo com uma energia incrível e contagiante. Rimos muito e vivemos muitas aventuras.

Agradeço também a todos que me apoiaram nas viagens de reportagem, especialmente Stephen Thuo Kiguru, meu intrépido guia em Nairóbi, sempre disponível para tudo o que eu precisava e cujo humor e otimismo incansáveis permaneceram intactos, mesmo quando alguém chamou a polícia devido a um grave mal-entendido após uma de nossas excursões. Esta é uma história para outra ocasião.

Aos meus queridos amigos e mentores: Angela Chen, Gideon Lichfield, Roger McNamee, Brenda Guadalupe López Alatorre, Jose Manuel Rodriguez Moreno, Bina Venkataraman e Tate Ryan-Mosley, que generosamente leram os primeiros rascunhos do livro ou diversos trechos e me ofereceram um feedback sábio e inestimável. A Oren Etzioni, que revisou com gentileza meu relato da história da IA e todas as minhas explicações técnicas sobre a pesquisa no assunto, garantindo que estivessem corretas e devidamente contextualizadas. E a Ria Kalluri, cuja amizade foi uma fonte de força e alegria muito antes de começarmos a investigar a natureza colonizadora do desenvolvimento da inteligência artificial, e cuja clareza intelectual e moral sobre o tema tem sido uma luz orientadora.

Este livro também se apoia em reportagens e trabalhos que realizei ao longo de minha carreira jornalística. Seria negligente de minha parte não agradecer a todas as pessoas que me apoiaram ao longo desse caminho. Toda a gratidão a Janet Guyon, minha editora na *Quartz*, que foi a primeira pessoa a perceber o que eu estava tentando dizer, e a me dizer que acreditava que eu poderia ser uma ótima jornalista. A Gideon, então editor-chefe da *MIT Technology Review*, que fez uma aposta ousada ao me oferecer meu primeiro emprego em tempo integral no jornalismo, para cobrir nada menos que inteligência artificial, colocando-me numa trajetória de anos que eu jamais poderia ter imaginado. A Niall Firth, meu editor na *Tech Review*, que um dia me perguntou: "Por que você não faz um perfil da OpenAI?", quando eu nunca tinha feito algo parecido de qualquer empresa. Por alguma razão, Niall acreditava que eu seria capaz. E eu trabalhei mais e me esforcei ainda mais para provar que ele estava certo.

Quando comecei a perceber, e a direcionar minha atenção, para a vasta desigualdade global que a IA estava perpetuando, Niall também foi fundamental ao apoiar a nova linha de investigação, assim como Angela, a colega que mencionei no capítulo 4, que identificou a expressão "colonialismo de dados" na literatura acadêmica e me ajudou a seguir na direção certa. Mat Honan, que substituiu Gideon como editor-chefe, compreendeu depressa a importância de minhas investigações e me apoiou incondicionalmente para que eu pudesse levá-las adiante.

No fim de 2021, tirei uma licença da *Tech Review* para me dedicar a um projeto de reportagem de seis meses sobre o "colonialismo da IA", com o maravilhoso auxílio da bolsa Knight Science Journalism do MIT e de uma bolsa do Pulitzer Center AI Accountability. Sou grata a Deborah Blum e Ashley Smart, do MIT KSJ, e a Marina Walker Guevara e Boyoung Lim, do Pulitzer Center, por me proporcionarem os recursos necessários para realizar um projeto tão amplo em plena pandemia. O resultado, uma série em quatro partes com histórias da África do Sul, Venezuela, Indonésia e Nova Zelândia, lançou as bases para a tese e, por fim, para o título deste livro. Agradeço às minhas incríveis colaboradoras nessas reportagens: Heidi Swart, Andrea Paola Hernández e Nadine Freischlad, cuja experiência jornalística, domínio linguístico e profundo conhecimento dos contextos locais e culturais tornaram as histórias mais ricas do que eu jamais poderia alcançar sozinha. Foi sob a liderança visionária de Marina em colaborações jornalísticas globais que me conectei com Heidi, Andrea e Nadine e aprendi uma nova abordagem para conduzir projetos de reportagem de alcance internacional. Desde então, a Rede de Responsabilidade em IA do Pulitzer Center, que reúne jornalistas do mundo inteiro para promover investigações sobre responsabilidade em inteligência artificial, tornou-se uma de minhas comunidades profissionais mais importantes.

No *Wall Street Journal*, foi meu editor Josh Chin quem primeiro me incentivou a apresentar o que acabaria se tornando uma reportagem de

primeira página sobre Mophat Okinyi e os trabalhadores quenianos contratados pela OpenAI; Drew Dowell e Jason Dean ajudaram a viabilizar a viagem para que isso fosse possível. No *The Atlantic*, meu editor Damon Beres não precisou ser convencido quando lhe propus uma matéria complexa sobre os impactos ambientais da infraestrutura computacional por trás da IA, nem Paul Bisceglio ou Adrienne LaFrance, que aprovaram outra viagem para o Arizona. Sou igualmente grata a Bradley Olson, Deepa Seetharaman, Daniel Engber e Matteo Wong por me ajudarem a concretizar essas matérias. Trabalhar ao lado de meus colegas inigualáveis em ambas as publicações me deu uma compreensão totalmente nova do que significa apurar e escrever matérias com excelência.

Por fim, expresso o amor mais profundo e gratidão à minha família. À minha mãe, que reconheceu meu amor pela escrita quando eu ainda era criança e dedicou tudo o que tinha para me ajudar a realizar meus sonhos. Ao meu pai, que jamais hesitou em fazer tudo o que podia para me apoiar. À minha 奶奶, minha fonte inesgotável de inspiração. Aos meus sogros, que sabiamente me lembram a importância de saborear o processo e celebrar cada vitória. E ao meu marido: melhor amigo, parceiro de vida, bússola moral, torcedor, fã número um, leitor precoce, porto seguro, conselheiro e romântico incansável. Amar você e ser amada por você é o alicerce de tudo o que faço.

# NOTAS

## EPÍGRAFES

1 Joseph Weizenbaum, "ELIZA — a Computer Program for the Study of Natural Language Communication Between Man and Machine", *Communications of the ACM* 9, n. 1, janeiro de 1966, p. 36-45, doi.org/10.1145/365153.365168.

2 Sam Altman, "Successful People", *Sam Altman* (blog), 7 de março de 2013, blog.samaltman.com/successful-people.

## PRÓLOGO: UMA CORRIDA PELO TRONO

1 Tripp Mickle, Cade Metz, Mike Isaac e Karen Weise, "Inside OpenAI's Crisis over the Future of Artificial Intelligence", *New York Times*, 9 de dezembro de 2023, nytimes.com/2023/12/09/technology/openai-altman-inside-crisis.html.

2 Trevor Noah, apresentador, *What Now? with Trevor Noah*, podcast, temporada 1, episódio 5, "Sam Altman Speaks Out about What Happened at OpenAI", Spotify Podcasts, 7 de dezembro de 2023, open.spotify.com/show/122imavATqSE7eCyXIcqZL.

3 OpenAI, "OpenAI Announces Leadership Transition", *OpenAI* (blog), 17 de novembro de 2023, openai.com/index/openai-announces-leadership-transition.

4 Com exceção de indicação que refute, os relatos internos das experiências de funcionários, diretores do conselho e da liderança durante a crise do conselho são de onze pessoas que estiveram presentes nos momentos descritos.

5 Todos os trechos de diálogo da reunião geral foram retirados de uma gravação de áudio da reunião, 17 de novembro de 2023.

6 Captura de tela do alerta, 17 de novembro de 2023.

7 Hannah Miller, Brad Stone, Shirin Ghaffary e Ashlee Vance, "Silicon Valley Boardroom Coup Leads to Ouster of an AI Champion", *Bloomberg*, 17 de novembro de 2023, bloomberg.com/news/articles/2023-11-18/openai-altman-ouster-followed-debates-between-altman-board.

8 Keach Hagey, Deepa Seetharaman e Berber Jin, "Behind the Scenes of Sam Altman's Showdown at OpenAI", *Wall Street Journal*, 22 de novembro de 2023, wsj.com/tech/ai/altman-firing-openai-520a3a8c.

9 Kate Clark, Natasha Mascarenhas e Anissa Gardizy, "If Sam Altman Returns to OpenAI, Board Will Go", *The Information*, 18 de novembro de 2023, theinformation.com/articles/altman-decision-looms-as-sequoia-tiger-negotiate-behind-scenes.

10 Erin Woo, Anissa Gardizy e Amir Efrati, "OpenAI 'Optimistic' It Can Bring Back Sam Altman, Greg Brockman", *The Information*, 18 de novembro de 2023, theinformation.com/articles/openai-optimistic-it-can-bring-back-sam-altman-greg-brockman?rc=ot38so.

11 Alex Konrad e David Jeans, "OpenAI Investors Plot Last-Minute Push with Microsoft to Reinstate Sam Altman as CEO", *Forbes*, 18 de novembro de 2023, forbes.com/sites/alexkonrad/2023/11/18/openai-investors-scramble-to-reinstate-sam-altman-as-ceo.

12 Todas as citações do Slack da OpenAI foram extraídas de capturas de tela.

13 Deepa Seetharaman, Berber Jin e Keach Hagey, "OpenAI Investors Keep Pushing for Sam Altman's Return", *Wall Street Journal*, 21 de novembro de 2023, wsj.com/tech/openai-employees-threaten-to-quit-unless-board-resigns-bbd5cc86.

14 Fotos da montagem no escritório.

15 Amir Efrati, Anissa Gardizy e Erin Woo, "Altman Agrees to Internal Investigation upon Return to OpenAI", *The Information*, 21 de novembro de 2023, theinformation.com/articles/breaking-sam-altman-to-return-as-openai-ceo.

16 Greg Brockman (@gdb), "estamos mais do que de volta", Twitter (atual X), 21 de novembro de 2024, x.com/gdb/

status/1727230819226583113.

17 Raffaele Huang e Karen Hao, "Baidu Hurries to Ready China's First ChatGPT Equivalent Ahead of Launch", *Wall Street Journal*, 9 de março de 2023, wsj.com/articles/baidu-scrambles-to-ready-chinas-first-chatgpt-equivalent-ahead-of-launch-bf359ca4.

18 Parmy Olson e Carolyn Silverman, "ChatGPT's $8 Trillion Birthday Gift to Big Tech", *Bloomberg*, 29 de novembro de 2024, bloomberg.com/opinion/articles/2024-11-29/chatgpt-turns-2-and-gives-8-trillion-birthday-gift-to-big-tech.

19 Goldman Sachs, *Gen AI: Too Much Spend, Too Little Benefit?*, 27 de junho de 2024, goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit.

20 "Upwork Study Finds Employee Workloads Rising Despite Increased C-Suite Investment in Artificial Intelligence", *Upwork*, 23 de julho de 2024, https://investors.upwork.com/news-releases/news-release-details/upwork-study-finds-employee-workloads-rising-despite-increased-c.

21 Olson e Silverman, "ChatGPT's $8 Trillion Birthday Gift".

22 Sam Altman, "The Intelligence Age", *Sam Altman* (blog), 23 de setembro de 2024, ia.samaltman.com.

## CAPÍTULO 1: DIREITO DIVINO

1 Cade Metz, *Criadores de gênios: os inovadores que levaram a IA para o Google, o Facebook e o mundo* (Alta Books, 2022).

2 Diversos relatos desta reunião foram publicados ao longo dos anos, inclusive em *Criadores de gênios: os inovadores que levaram a IA para o Google, o Facebook e o mundo.* Greg Brockman também descreveu sua própria versão em dois posts de blog: Greg Brockman, "My Path to OpenAI", *Greg Brockman* (blog), 3 de maio de 2016, blog.gregbrockman.com/my-path-to-openai; e Greg Brockman, "#define CTO OpenAI", *Greg Brockman* (blog), 9 de janeiro de 2017, blog.gregbrockman.com/define-cto-openai.

3 As opiniões de Musk sobre Altman, sua experiência como cofundador da OpenAI e a evolução de suas visões sobre a IA baseiam-se, em grande parte, em um processo movido

por Musk contra Altman, Brockman e a OpenAI em 29 de fevereiro de 2024, reaberto em 5 de agosto de 2024: *Musk v. Altman*, nº 4:24-cv-04722, *CourtListener* (N. D. Cal., 5 de agosto de 2024). Informações adicionais vêm principalmente de Maureen Dowd, "Elon Musk's Future Shock", *Vanity Fair*, abril de 2017, archive.vanityfair.com/article/2017/4/elon-musks-future-shock; e Walter Isaacson, *Elon Musk* (Intrínseca, 2023).

4 Lex Fridman, apresentador, *Lex Fridman Podcast*, podcast, episódio 367, "Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI", 25 de março de 2023, lexfridman.com/podcast.

5 Sam Altman, "How to Be Successful", *Sam Altman* (blog), 24 de janeiro de 2019, blog.samaltman.com/how-to-be-successful.

6 Entrevista da autora com Timnit Gebru, agosto de 2023.

7 Tad Friend, "Sam Altman's Manifest Destiny", *New Yorker*, 3 de outubro de 2016, newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny.

8 Isaacson, *Elon Musk*.

9 "Decoding Google Gemini with Jeff Dean", publicado em 11 de setembro de 2024 pela Google DeepMind, YouTube, 55 min 55 s, youtu.be/lH74gNeryhQ; correspondência do autor com porta-voz do Google, novembro de 2024.

10 Com base nas lembranças e descrições de quatro pessoas que conversaram com Musk ou estavam presentes quando ele expressou suas opiniões, bem como em *Musk v. Altman*, *CourtListener*, ECF nº 32, Anexo 13.

11 Um porta-voz da Google DeepMind também rejeitou a descrição de Hassabis feita por Musk. Correspondência do autor com porta-voz da Google DeepMind, novembro de 2024.

12 Nick Bostrom, *Superinteligência: caminhos, perigos, estratégias* (Darkside Books, 2018).

13 Elon Musk (@elonmusk), "Vale a pena ler Superinteligência de Bostrom. Precisamos ser supercuidadosos com a IA. Potencialmente mais perigosa que armas nucleares", Twitter (atual X), 3 de agosto de 2014, x.com/elonmusk/status/495759307346952192.

14 Nick Bostrom, "Apology for an Old Email", *Nick Bostrom's Home Page*, 9 de janeiro de 2023, nickbostrom.com/oldemail.pdf.

15 Olivia Carville, "The Super Rich of Silicon Valley Have a Doomsday Escape Plan in New Zealand", *Bloomberg*, 5 de setembro de 2018, bloomberg.com/features/2018-rich-new-zealand-doomsday-preppers.

16 Sam Altman, "Machine Intelligence, Part 1", *Sam Altman* (blog), 25 de fevereiro de 2015, blog.samaltman.com/machine-intelligence-part-1.

17 Toda a correspondência por e-mail entre Musk e Altman neste capítulo vem do processo judicial movido por Musk, anexada como prova ao documento número 32: *Musk,* CourtListener, ECF nº 32.

18 Melia Russell e Julia Black, "Ele jogou xadrez com Peter Thiel, discutiu com Elon Musk e, certa vez, supostamente impediu a queda de um avião: dentro do mundo de Sam Altman, onde a verdade é mais estranha que a ficção", *Business Insider*, 27 de abril de 2023, businessinsider.com/sam-altman-openai-chatgpt-worldcoin-helion-future-tech-2023-4.

19 Paul Graham, "A Fundraising Survival Guide", *Paul Graham* (blog), acesso em 21 de novembro de 2024, paulgraham.com/fundraising.html.

20 Friend, "Sam Altman's Manifest Destiny".

21 "Megan O'Neill Is Wed to Jerold D. Altman", *New York Times*, 24 de julho de 1977, nytimes.com/1977/07/24/archives/megan-oneill-is-wed-to-jerold-d-altman.html.

22 O relato de infância de Altman baseia-se, grande parte, em três perfis principais sobre ele: Friend, "Sam Altman's Manifest Destiny"; Elizabeth Weil, "Sam Altman Is the Oppenheimer of Our Age", *New York*, 25 de setembro de 2023, nymag.com/intelligencer/article/sam-altman-artificial-intelligence-openai-profile.html; e Ellen Huet, apresentadora, *Foundering: The OpenAI Story*, podcast, temporada 5, episódio 1, "The Most Silicon Valley Man Alive", Bloomberg Podcasts, 5 de junho de 2024, https://www.bloomberg.com/news/audio/2024-06-05/foundering-openai-e1-most-silicon-valley-man-alive-podcast

23 "Sam Altman: How to Build the Future", publicado em 27 de setembro de 2016 pela Y Combinator, YouTube, 20 min 9 s, youtu.be/sYMqVwsewSg.

24 Huet, "The Most Silicon Valley Man Alive".

25 Parmy Olson, *Supremacy: AI, ChatGPT, and the Race that Will Change the World* (St. Martin's Press, 2024), 5.

26 Weil, "Sam Altman Is the Oppenheimer of Our Age".

27 Friend, "Sam Altman's Manifest Destiny".

28 Joe Hudson e Brett Kistler, apresentadores, *The Art of Accomplishment Podcast*, podcast, episódio 39, "Sam Altman — Leading with Crippling Anxiety, Discovering Meditation, and Building Intelligence with Self-Awareness", 14 de janeiro de 2022, artofaccomplishment.com/podcast.

29 Friend, "Sam Altman's Manifest Destiny."

30 "Office Hours with Sam Altman", publicado em 11 de janeiro de 2017 pela Y Combinator, YouTube, 24 min 34 s, youtu.be/45BvnJgwYjk.

31 Russell e Black, "He's Played Chess with Peter Thiel".

32 Deepa Seetharaman, Keach Hagey, Berber Jin e Kate Linebaugh, "Sam Altman's Knack for Dodging Bullets — with a Little Help from Bigshot Friends", *The Wall Street Journal*, 24 de dezembro de 2023, wsj.com/tech/ai/sam-altman-openai-protected-by-silicon-valley-friends-f3efcf68.

33 "Sam Altman Startup School Video", publicado em 26 de julho de 2017 por Waterloo Engineering, YouTube, 1 h 18 min 19 s, youtu.be/4SlNgM4PjvQ.

34 "Paper Chase", *Venture Capital Journal*, 1º de dezembro de 2006, venturecapitaljournal.com/paper-chase.

35 Annie Massa e Vernal Galpotthawela, "Sam Altman Is Worth $2 Billion — That Doesn't Include OpenAI", *Bloomberg*, 1º de março de 2024, bloomberg.com/news/articles/2024-03-01/sam-altman-is-a-billionaire-thanks-to-vc-funds-startups.

36 "First Look: Loopt Provides More Incentives to Try Location-Based Services with Loopt Star", publicado em 31 de maio de 2010 por Robert Scoble, YouTube, 15 min 24 s, youtu.be/P5izvkusAMM.

37 "Why Loopt Partnered with Facebook", publicado em 3 de novembro de 2010 por CNN Business, YouTube, 2 min 41 s, youtu.be/tMO0Gm6yxWc.

38 Jessica E. Lessin, "This Is How Sam Altman Works the Press and Congress. I Know from Experience", *The Information*, 7 de junho de 2023, theinformation.com/articles/this-is-how-sam-altman-works-the-press-and-congress-i-know-from-experience.

39 Russell e Black, "He's Played Chess with Peter Thiel".

40 Christine Lagorio-Chafkin, "Inside Reddit's Long, Complicated Relationship with OpenAI's Sam Altman", *Inc.*, 8 de março de 2024, inc.com/christine-lagorio/inside-reddits-long-complicated-relationship-with-openais-sam-altman.html.

41 Entrevista da autora com Geoff Ralston, março de 2024.

42 Seetharaman et al., "Sam Altman's Knack for Dodging Bullets".

43 Walter Isaacson, *Steve Jobs* (Simon & Schuster, 2011), 104.

44 Katie Notopoulos, "Sam Altman Is Seen Driving a Car That Can Cost $5 Million. Everyone Is Thanking Him for Helping Them Pass Their Tests", *Business Insider*, 12 de julho de 2024, https://africa.businessinsider.com/news/sam-altman-is-seen-driving-a-car-that-can-cost-dollar5-million-everyone-is-thanking/6tqddlp.

45 Elizabeth Dwoskin, Marc Fisher e Nitasha Tiku, "'King of the Cannibals': How Sam Altman Took Over Silicon Valley", *Washington Post*, 23 de dezembro de 2023, washingtonpost.com/technology/2023/12/23/sam-altman-openai-peter-thiel-silicon-valley.

46 Eric Newcomer, "YC's Paul Graham: The Complete Interview", 26 de dezembro de 2013, *The Information*, theinformation.com/articles/yc-s-paul-graham-the-complete-interview.

47 Paul Graham, "A Student's Guide to Startups", *Paul Graham* (blog), outubro de 2006, paulgraham.com/mit.html.

48 Paul Graham, "What We Look for in Founders", *Paul Graham* (blog), outubro de 2010, paulgraham.com/founders.html.

49 Paul Graham, “Five Founders”, *Paul Graham* (blog), abril de 2009, paulgraham.com/5founders.html.

50 Friend, “Sam Altman’s Manifest Destiny”.

51 Dwoskin et al., “‘King of the Cannibals’”.

52 Sam Altman, “Growth and Government”, *Sam Altman* (blog), 4 de março de 2013, blog.samaltman.com/growth-and-government.

53 “Sam Altman Startup School Video”, Waterloo Engineering.

54 Tyler Cowen, apresentador, *Conversations with Tyler*, podcast, episódio 61, “Sam Altman on Loving Community, Hating Coworking, and the Hunt for Talent”, Mercatus Center Podcasts, 27 de fevereiro de 2019.

55 “Competition Is for Losers with Peter Thiel (How to Start a Startup 2014: 5)”, publicado em 22 de março de 2017 pela Y Combinator, YouTube, 50 min. 27 s, youtu.be/3Fx5Q8xGU8k.

56 Sam Altman, “How Things Get Done”, *Sam Altman* (blog), 17 de julho de 2013, blog.samaltman.com/how-things-get-done.

57 “Sam Altman Startup School Video”, Waterloo Engineering.

58 Berber Jin, Tom Dotan e Keach Hagey, “The Opaque Investment Empire Making OpenAI’s Sam Altman Rich”, *Wall Street Journal*, 3 de junho de 2024, wsj.com/tech/ai/openai-sam-altman-investments-004fc785.

59 Entrevista da autora com Matt Krisiloff, abril de 2024.

60 Sam Altman, “The United Slate”, *Sam Altman* (blog), 12 de julho de 2017, blog.samaltman.com/the-united-slate.

61 Dwoskin et al., “‘King of the Cannibals’”.

62 Friend, “Sam Altman’s Manifest Destiny”.

63 Uma foto da visita de Schumer ao Pioneer Building, 8 de março de 2019.

64 Entrevistas da autora com Annie Altman, de março a novembro de 2024.

65 Sam Altman (@sama), “Minha irmã entrou com um processo contra mim. Aqui está uma declaração minha, de minha mãe e de meus irmãos:”, Twitter (atual X), 7 de janeiro de 2025, x.com/sama/status/1876780763653263770.

66 Correspondência da autora com Connie Gibstine, outubro de 2024.

67 Altman v. Altman, No. 4:25-cv-00017, CourtListener (E.D. Mo. 06 de janeiro de 2025) ECF No. 1.

## CAPÍTULO 2: UMA MISSÃO CIVILIZADORA

1 Entrevistas da autora com Greg Brockman, agosto de 2019.

2 A história da formação e dos primeiros anos de Sutskever baseia-se parcialmente em diversas entrevistas concedidas à imprensa, incluindo: "Interview with Dr. Ilya Sutskever, Co-founder of OPEN AI — at the Open University Studios — English", publicado em 13 de setembro de 2023 pela The Open University of Israel, YouTube, 50 min 28 s, youtu.be/H1YoNlz2LxA; Nina Haikara, "This U of T Alum Is Leading AI Research at $1 Billion Non-Profit Backed by Elon Musk", *U of T News*, 28 de março de 2017, utoronto.ca/news /u-t-alum-leading-ai-research-1-billion-non-profit-backed-elon-musk; e Varsity Contributor, "Neural Networking", *The Varsity*, 25 de outubro de 2010, thevarsity.ca/2010/10/25/neural-networking.

3 Os resultados revolucionários, obtidos em 2012, foram publicados em um periódico cinco anos depois: Alex Krizhevsky, Ilya Sutskever e Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Communications of the ACM* 60, n. 6, maio de 2017, p. 84-90, doi.org/10.1145/3065386.

4 Entrevista da autora com Geoff Hinton, agosto de 2023.

5 Cade Metz, *Criadores de gênios: os inovadores que levaram a IA para o Google, o Facebook e o Mundo* (Alta Books, 2022).

6 Greg Brockman, "#define CTO OpenAI", *Greg Brockman* (blog), 9 de janeiro de 2017, blog.gregbrockman.com/define-cto-openai.

7 Sam Altman, "Greg", *Sam Altman* (blog), 7 de março de 2017, blog.samaltman.com/greg.

8 Entrevista da autora com Pieter Abbeel, agosto de 2019.

9 Metz, *Criadores de gênios*.

10 Toda a correspondência entre as lideranças da OpenAI e da Tesla neste capítulo provém do

processo movido por Musk, anexada como provas ao documento n. 32: *Musk v. Altman*, No. 4:24-cv-04722, *CourtListener* (N. D. Cal., 14 de novembro de 2024), ECF n. 32; e das respostas da OpenAI publicadas no blog da empresa: OpenAI, "OpenAI and Elon Musk", *OpenAI* (blog), 5 de março de 2024, openai.com/index/openai-elon-musk; OpenAI, "Elon Musk Wanted an OpenAI For-Profit", *OpenAI* (blog), 13 de dezembro de 2024, openai.com/index/elon-musk-wanted-an-openai-for-profit/#summer-elon-that-a-for-profit-was-the-next-step-for-mission.

11 *Musk, CourtListener*, ECF n. 32, Anexo 7.

12 "OpenAI Inc", *ProPublica Nonprofit Explorer*, acesso em 25 de agosto de 2024, projects.propublica.org/nonprofits/organizations/810861541/201703459349300445/full.

13 Metz, *Criadores de gênios.*

14 Metz, *Criadores de gênios.*

15 *Musk, CourtListener*, ECF n. 32, Anexo 7.

16 Walter Isaacson, *Elon Musk* (Intrínseca, 2023).

17 Uma contribuição seminal para o entendimento de como a IA leva à discriminação vem de Solon Barocas e Andrew D. Selbst, "Big Data's Disparate Impact", *Califórnia Law Review*, 104, n. 3, 2016, p. 671-732, ssrn.com/abstract=2477899. Também há uma reportagem que explora mais detalhadamente como essa discriminação se manifesta na prática: Karen Hao, "The Coming War on the Hidden Algorithms That Trap People in Poverty", *MIT Technology Review*, 4 de dezembro de 2020, technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back.

18 Alexandra Stevenson, "Facebook Admits It Was Used to Incite Violence in Myanmar", *New York Times*, 6 de novembro de 2018, nytimes.com/2018/11/06/technology/myanmar-facebook.html.

19 Eric Lipton, "As A.I.-Controlled Killer Drones Become Reality, Nations Debate Limits", *New York Times*, 21 de novembro de 2023, nytimes.com/2023/11/21/us/politics/ai-drones-war-law.html.

20 Entrevista da autora com Chip Huyen, agosto de 2019.

21 Entrevista da autora com Timnit Gebru, março de 2021.

22 Cópia do e-mail fornecida por Gebru.

23 Entrevista da autora com Greg Brockman, agosto de 2019.

24 Entrevista com Brockman.

25 Arthur C. Clarke, *Profiles of the Future: An Inquiry into the Limits of the Possible* (Bantam Books, 1962), p. 30-39.

26 Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman e Dan Mané, "Concrete Problems in AI Safety", preprint, *arXiv*, 25 de julho de 2016, p. 1-29, doi.org/10.48550/arXiv.1606.06565.

27 Correspondência da autora com porta-voz da Open Philanthropy, novembro de 2024.

28 Julia Angwin, Jeff Larson, Surya Mattu e Lauren Kirchner, "Machine Bias", *ProPublica*, 23 de maio de 2016, propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

29 O artigo foi apresentado na *International Conference of Learning Representations* no workshop "Machine Learning in Real Life", em 26 de abril de 2020, sites.google.com/nyu.edu/ml-irl-2020/home, e publicado posteriormente no arXiv: Inioluwa Deborah Raji e Roel Dobbe, "Concrete Problems in AI Safety, Revisited", *arXiv*, 18 de dezembro de 2023, p. 1-6, doi.org/10.48550/arXiv.2401.10899.

30 Tad Friend, "Sam Altman's Manifest Destiny", *The New Yorker*, 3 de outubro de 2016, newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny.

31 "OpenAI — General Support", *Open Philanthropy*, acesso em 27 de novembro de 2024, openphilanthropy.org/grants/openai-general-support.

32 Entrevista da autora com Greg Brockman e Daniela Amodei, agosto de 2019.

33 *ProPublica Nonprofit Explorer*, "OpenAI Inc".

34 Entrevista da autora com Greg Brockman e Ilya Sutskever, agosto de 2019.

35 Entrevista com Brockman e Sutskever.

36 OpenAI, "AI and Compute", *OpenAI* (blog), 16 de maio de 2018, openai.com/index/ai-

and-compute.

37 *Musk, CourtListener*, ECF n. 32; Id., ECF nº 32, Anexo 11.

38 Id., ECF n. 32, Anexo 13; OpenAI, "OpenAI and Elon Musk"; OpenAI, "Elon Musk Wanted an OpenAI For-Profit".

39 Entrevistas com Brockman, agosto de 2019.

40 OpenAI, "OpenAI and Elon Musk".

41 *Musk, CourtListener*, ECF n. 32, Anexo 15.

42 Id., ECF n.º32, Anexo 7.

43 Berber Jin e Keach Hagey, "The Contradictions of Sam Altman, AI Crusader Behind ChatGPT", *Wall Street Journal*, 31 de março de 2023, wsj.com/tech/ai/chatgpt-sam-altman-artificial-intelligence-openai-b0e1c8c9.

44 Correspondência com Jennifer 8. Lee, coprodutora do documentário *Artificial Gamer*, dirigido por Chad Herschberger e com participação de Pieter Abbeel, Greg Brockman e Noam Brown, lançado em 24 de setembro de 2021, artificialgamerfilm.com. A equipe do documentário manteve independência de edição sobre o filme.

45 OpenAI, "OpenAI Charter", *OpenAI* (blog), acessado em 25 de agosto de 2024, openai.com/charter.

46 Jin e Hagey, "The Contradictions of Sam Altman".

47 Entrevista da autora com Xuedong Huang, julho de 2023.

48 *Musk, CourtListener*, ECF n. 1, *5.

49 Elizabeth Dwoskin e Nitasha Tiku, "Altman's Polarizing Past Hints at OpenAI Board's Reason for Firing Him", *Washington Post*, 22 de novembro de 2023, washingtonpost.com/technology/2023/11/22/sam-altman-fired-y-combinator-paul-graham.

50 Deepa Seetharaman, Keach Hagey e Berber Jin, "Sam Altman's Knack for Dodging Bullets — with a Little Help from Bigshot Friends", *Wall Street Journal*, 24 de dezembro de 2023, wsj.com/tech/ai/sam-altman-openai-protected-by-silicon-valley-friends-f3efcf68.

51 Cópia do documento sobre a estrutura de faixas salariais (*payband document*).

52 Karen Hao, "The Messy, Secretive Reality

Behind OpenAI's Bid to Save the World", *MIT Technology Review*, 17 de fevereiro de 2020, technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality.

53 @windowshopping, "Eu estava acreditando até ele dizer que o lucro é 'limitado' a 100 vezes o investimento inicial. Então alguém que investe 10 milhões de dólares tem seu retorno 'limitado' a um bilhão. Rá, Rá. Basicamente ilimitado, a menos que a empresa cresça ao nível de mercado das FAANG", *Hacker News*, 11 de março de 2019, news.ycombinator.com/item?id=19360709.

54 Todos os valores numéricos referentes aos investimentos no fundo limitado (LP) e seus limites de lucro mencionados ao longo do livro vêm de um documento financeiro interno da OpenAI.

55 Chamath Palihapitiya, Jason Calacanis, David Sacks e David Friedberg, apresentadores, *All-In,* podcast, episódio 194, "In Conversation with Reid Hoffman & Robert F. Kennedy Jr.", 30 de agosto de 2024, https://allin.com/episodes.

56 Os e-mails de Kevin Scott e Satya Nadella foram divulgados em 2024 como parte do processo antitruste do Departamento de Justiça dos EUA contra o Google. Jyoti Mann e Beatrice Nolan, "Read the Email to Satya Nadella and Bill Gates That Shows Microsoft's CTO Was 'Very Worried' about Google's AI Progress in 2019", *Business Insider*, 1º de maio de 2024, businessinsider.com/satya-nadella-bill-gates-microsoft-concern-google-rivals-ai-emails-2024-5.

## CAPÍTULO 3: CENTRO NERVOSO

1 Eddie Sun, "ChatGPT's San Francisco Offices Getting Nap Rooms, a Museum for Staffers", *San Francisco Standard*, 11 de julho de 2023, https://sfstandard.com/2023/07/11/chatgpt-secretive-san-francisco-offices-nap-rooms-museum-open-ai.

2 Eu estimei o preço dos móveis fazendo buscas reversas de imagem, usando o Google Images, com fotos do escritório. Quando

tentei obter o preço exato com o escritório de arquitetura que trabalhou para a OpenAI no Mayo, assim como confirmar minhas descrições ("É correto dizer que a escada em espiral do escritório é feita de madeira e pedra?"), uma pessoa respondeu que a empresa assinou um acordo de confidencialidade e não pode falar sobre o projeto.

3 Berber Jin e Keach Hagey, "The Contradictions of Sam Altman, AI Crusader", *Wall Street Journal*, 31 de março de 2023, wsj.com/tech/ai/chatgpt-sam-altman-artificial-intelligence-openai-b0e1c8c9.

4 Entrevista da autora com Ben Barry, outubro de 2023.

5 Neurips, "NeurIPS 2019 Workshop: Tackling Climate Change with Machine Learning" ,Vancouver Convention Center, 14 de dezembro de 2019, https://www.climatechange.ai/events/neurips2019; e ML4H, "Machine Learning for Health – NeurIPS Workshop", Vancouver Convention Center, 13 de dezembro de 2019, https://ml4h.cc/2019/index.html.

6 David Rolnick et al., "Tackling Climate Change with Machine Learning", *ACM Computing Surveys*, v. 55, n. 2, fevereiro de 2022, p. 1-96, https://doi.org/10.1145/3485128.

7 Emma Strubell, Ananya Ganesh e Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, , 2019, p. 3645-3650, https://doi.org/10.18653/v1/P19-1355.

8 Cade Metz, *Criadores de gênios: os inovadores que levaram a IA para o Google, o Facebook e o mundo* (Alta Books, 2022).

9 Entrevista da autora com Sam Brockman, agosto de 2019.

10 Karen Hao, "The Messy, Secretive Reality Behind OpenAI's Bid to Save the World", *MIT Technology Review*, 17 de fevereiro de 2020, https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-greg-messy-secretive-reality.

11 Elon Musk (@elonmusk), "A OpenAI

deveria ser mais aberta, em minha opinião", Twitter (atual X), 17 de fevereiro de 2020, https://x.com/elonmusk/status/1229544673590599681.

12 Sam Altman. Cópia do e-mail enviado à OpenAI.

## CAPÍTULO 4: SONHOS DE MODERNIDADE

1 Daron Acemoglu e Simon Johnson, *Poder e progresso: uma luta de mil anos entre a tecnologia e a prosperidade* (Objetiva, 2024).

2 John McCarthy, Marvin L. Minsky, Nathaniel Rochester e Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence", Universidade Stanford, 31 de agosto de 1955, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

3 No início, John McCarthy, Claude Shannon e outros reuniram artigos em um compêndio sobre o mesmo conjunto de ideias que mais tarde seria chamado de IA, intitulado *Automata Studies* [Estudos sobre autômatos], publicado em 1956. McCarthy ficou decepcionado com os artigos submetidos e com a falta de ambição deles. Ele afirmou que foi essa decepção que o levou a começar a usar o termo "inteligência artificial".

4 Kate Crawford, *Atlas da IA: poder, política e os custos planetários da inteligência artificial* (Edições Sesc, 2025).

5 John McCarthy, "What Is Artificial Intelligence?", *Formal Reasoning Group*, 12 de novembro de 2007, http://www-formal.stanford.edu/jmc/whatisai.pdf.

6 Jenna Burrel, "Artificial Intelligence and the Ever-Receding Horizon of the Future", *Tech Policy Press*, 6 de junho de 2023, https://techpolicy.press/artificial-intelligence-and-the-ever-receding-horizon-of-the-future.

7 Marvin Minsky e Seymour A. Papert, *Perceptrons: An Introduction to Computational Geometry*, (MIT Press, 1969).

8 Joseph Weizenbaum, "ELIZA: A computer program for the study of natural language communication between man and machine", *Communications of the ACM*, v. 9, n. 1, janeiro de 1966, p. 36-45, https://doi.org/10.1145/365153.365168.

9 Joseph Weizenbaum, "ELIZA: A computer

program for the study of natural language communication between man and machine", *Communications of the ACM*, v. 9, n. 1, janeiro de 1966, p. 36-45, https://doi.org/10.1145/365153.365168.

10 Ben Tarnoff, "Weizenbaum's Nightmares: How the Inventor of the First Chatbot Turned Against AI", *The Guardian*, 25 de julho de 2023, https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai.

11 Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (W. H. Freeman & Co, 1976).

12 Não existe uma história única sobre quando ocorreu cada inverno da IA. De modo geral, considera-se que o primeiro aconteceu nos anos 1970, desencadeado por um relatório de 1973 do British Science Research Council, elaborado pelo professor Sir James Lighthill, da Universidade de Cambridge, intitulado "Artificial Intelligence: A General Survey" [Inteligência artificial: uma visão geral]. Nele, Lighthill observou: "Em nenhuma área do campo da IA as descobertas feitas até agora produziram o impacto significativo que havia sido prometido na época." O segundo inverno da IA teria ocorrido aproximadamente entre o fim da década de 1980 e o início da década de 1990. Alguns estudiosos argumentam que, durante esse período, embora o financiamento para pesquisas rotuladas como "IA" tenha secado, o dinheiro continuou a ser direcionado ao desenvolvimento de técnicas relevantes sob outros nomes. Alguns pesquisadores também apontam para outros invernos da inteligência artificial. A professora de Stanford e referência na área, Fei-Fei Li, descreve o fim dos anos 1990 como um terceiro inverno em seu livro *The Worlds I See: Curiosity, Exploration, and Discovery at the Dawn of AI* (Flatiron Books, 2023), p. 89-90.

13 Entrevista da autora com Geoffrey Hinton, agosto de 2023.

14 Shoshana Zuboff, *A era do capitalismo de vigilância: a luta por um futuro humano na nova fronteira do poder* (Intrínseca, 2021).

15 Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini e

Abeba Birhane "The Surveillance AI Pipeline", preprint, *arXiv,* 17 de outubro de 2023, p. 10-11,https://doi.org/10.48550/arXiv.2309.15084.

16 Olivia Solon, "Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped Without Consent", *NBC News*, 12 de março de 2019, https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

17 Por exemplo, não demoraria para que começassem a surgir relatos de como o uso equivocado de sistemas defeituosos de reconhecimento facial estava levando à identificação incorreta de suspeitos e a prisões injustas. Até novembro de 2024, seis das sete pessoas reconhecidas que foram acusadas injustamente nos Estados Unidos, o que levou algumas à prisão, à perda do emprego, à separação de seus filhos e a relações pessoais destruídas, eram negras. Kashmir Hill, "Wrongfully Accused by an Algorithm", *New York Times*, 24 de junho de 2020, https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html; Khari Johnsson, "How Wrongful Arrests Based on AI Derailed 3 Men's Lives", *Wired*, 7 de março de 2022, https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives.

18 Max Newlon, "ISTE 2017: Most Innovative Winning Pitch", publicado em 13 de julho de 2018, YouTube, 7 min 42 s, https://youtu.be/oJt6cjdMGb4.

19 Jane Li, "A 'Brain-Reading' Headband Is Facing a Backlash in China", *Quartz*, 5 de novembro de 2019, https:/qz.com/1742279/a-mind-reading-headband-is-facing-backlash-in-china.

20 Nick Couldry e Ulises A. Mejias, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism* (Stanford University Press, 2019). Para ler mais sobre o conceito de extrativismo, ver: Rosemary Collard e Jessica Dempsey, "Extractivism Is Destroying Nature: To Tackle It Cop15 Must Go Beyond Simple Targets", *The Guardian*, 8 de dezembro de 2022, https://www.theguardian.com/environment/2022/dec/08/extractivism-is-destroying-nature-to-tackle-it-cop15-must-

go-beyond-simple-targets; e um dos principais textos sobre o conceito: Eduardo Gudynas, "Diez tesis urgentes sobre el nuevo extractivismo: Contextos y demandas bajo el progresismo sudamericano actual", em *Extractivismo, Política y Sociedad*, p. 187, 2009, https://www.rosalux.org.ec/pdfs/extractivismo.pdf.

21 Shakir Mohamed, Marie-Therese Png e William Isaac, "Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence", *Philosophy and Technology*, v. 33, p. 659-684, 12 de julho de 2020, https://doi.org/10.1007/s13347-020-00405-8.

22 Karen Hao e Heidi Swart, "South Africa's Private Surveillance Machine Is Fueling a Digital Apartheid", *MIT Technology Review*, 19 de abril de 2022, https://www.technologyreview.com/2022/04/19/1049996/south-africa-ai-surveillance-digital-apartheid.

23 Nestor Maslej et al., "AI Index Report 2024", p. 242, Stanford: Institute for Human-Centered AI, Universidade Stanford, abril de 2024, https://aiindex.stanford.edu/report.

24 Steven Rosenbush, "Big Tech Is Spending Billions on AI Research. Investors Should Keep an Eye Out", *The Wall Street Journal*, 8 de março de 2022, https://www.wsj.com/articles/big-tech-is-spending-billions-on-ai-research-investors-should-keep-an-eye-out-11646740800.

25 Nur Ahmed, Muntasir Wahed e Neil C. Thompson, "The growing influence of industry in AI research", *Science*, v. 379, n. 6635, 2 de março de 2023, p. 884-886, https://doi.org/10.1126/science.ade2420.

26 Nur Ahmed, Muntasir Wahed e Neil C. Thompson, "The growing influence of industry in AI research", *Science*, v. 379, n. 6635, 2 de março de 2023, p. 884-886, https://doi.org/10.1126/science.ade2420.

27 Em 2017, Tom Eck, diretor de tecnologia de plataformas industriais da IBM, disse de forma célebre: "Os pesquisadores de IA de mais alto nível estão recebendo salários de quarterbacks da NFL, o que já diz muito sobre a demanda e o valor percebido." Dan Butcher, "If you really know about artificial intelligence, you could earn as much as an NFL quarterback", *eFinancialCareers*, 13 de

julho de 2017, https://www.efinancialcareers.com/news/2017/07/top-talent-earns-high-ai-salaries-nfl-quarterbacks.

28 Mike Ramsey e Douglas MacMillan, "Carnegie Mellon Reels After Uber Lures Away Researchers", *The Wall Street Journal*, 31 de maio de 2015, https://www.wsj.com/articles/is-uber-a-friend-or-foe-of-carnegie-mellon-in-robotics-1433084582.

29 Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan e Michelle Bao. "The values encoded in machine learning research", em *AccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), p. 173-184, https://doi.org/10.1145/3531146.3533083.

30 The Outspoken Nomad, "Tesla FSD Beta: What-the-Hell Moments", publicado em 20 de janeiro de 2022, YouTube, 15 min 55 s, https://youtu.be/RVkLI9pPd24?t=166.

31 "Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018", Highway Accident Report, NTSB/HAR-19/03, PB2019-101402, *National Transportation Safety Board*, 19 de novembro de 2019, https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf.

32 "Additional Information Regarding EA22002", p. 1-6, 25 de abril de 2024, https://static.nhtsa.gov/odi/inv/2022/INCR-EA22002-14496.pdf.

33 Karen Hao, "Hackers Trick a Tesla into Veering into the Wrong Lane", *MIT Technology Review*, 1º de abril de 2019, https://www.technologyreview.com/2019/04/01/65915/hackers-trick-teslas-autopilot-into-veering-towards-oncoming-traffic.

34 Will Knight, "How Malevolent Machine Learning Could Derail AI", *MIT Technology Review*, 25 de março de 2019, https://www.technologyreview.com/2019/03/25/1216/emtech-digital-dawn-song-adversarial-machine-learning.

35 Benjamin Wilson, Judy Hoffman e Jamie Morgenstern, "Predictive inequity in object

detection", preprint, *arXiv*, 21 de fevereiro de 2019, p. 1-13, https://doi.org/10.48550/arXiv.1902.11097.

36 Xinyue Li, Zhenpeng Cheng, Jie M. Zhang, Federica Sarro, Ying Zhang e Xuanzhe Liu, "Bias behind the wheel: fairness analysis of autonomous driving systems", *ACM Transactions on Software Engineering and Methodology*, novembro de 2024, https://doi.org/10.1145/3702989.

37 Entrevista da autora com Deborah Raji, abril de 2020.

38 Entrevista da autora com Geoffrey Hinton, concedida durante o evento EmTech MIT, promovido pela *MIT Technology Review*, 20 de outubro de 2020. Para uma transcrição da entrevista, ver Karen Hao, "AI Pioneer Geoff Hinton: 'Deep Learning Is Going to Be Able to Do Everything'", *MIT Technology Review*, 3 de novembro de 2020, https://www.technologyreview.com/2020/11/03/1011616/ai-godfather-geoffrey-hinton-deep-learning-will-do-everything.

39 Entrevista da autora com Gary Marcus, setembro de 2019. Para uma transcrição da entrevista, ver: Karen Hao, "We Can't Trust AI Systems Built on Deep Learning Alone", *MIT Technology Review*, 27 de setembro de 2019, https://www.technologyreview.com/2019/09/27/65250/we-cant-trust-ai-systems-built-on-deep-learning-alone.

40 "Planning for AGI and Beyond", *OpenAI* (blog), 24 de fevereiro de 2023, https://openai.com/index/planning-for-agi-and-beyond.

41 Kevin Roose, "Bing's A.I. Chat: 'I Want to Be Alive'", *The New York Times*, 16 de fevereiro de 2023, https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

42 Pierre-François Lovens, "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là", *La Libre*, 28 de março de 2023, https://lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24.

43 Há vários artigos que encontraram esse resultado, incluindo um coescrito por Jacob

Hilton, que era pesquisador da OpenAI na época. Hilton e seus coautores descobriram que "os maiores modelos eram, em geral, os menos verídicos". Stephanie Lin, Jacob Hilton e Owain Evans, "TruthfulQA: measuring how models mimic human falsehoods", em *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, v. 1. 2021. p. 3214-3252, https://doi.org/10.18653/v1/2022.acl-long.229.
Além disso, para uma excelente explicação de por que os desenvolvedores têm se tornado cada vez menos conscientes da composição de seus dados de treinamento, ver: Christo Buschek e Jer Thorp, "Models All the Way Down", *Knowing Machines*, 26 de março de 2024, https://knowingmachines.org/models-all-the-way.

44 Benjamin Weiser, "Here's What Happens When Your Lawyer Uses ChatGPT", *The New York Times*, 27 de maio de 2023, https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

45 Katharina Jeblick et al., "ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports", *European Radiology*, v. 34, outubro de 2024, p. 2817-2825, https://doi.org/10.1007/s00330-023-10213-1.

46 Lily Hay Newman e Andy Greenberg, "Security News This Week: ChatGPT Spit Out Sensitive Data When Told to Repeat 'Poem', Forever", *Wired*, 2 de dezembro de 2023, https://www.wired.com/story/chatgpt-poem-forever-security-roundup.

47 Leonardo Nicoletti e Dina Bass, "Humans Are Biased. Generative AI Is Even Worse", *Bloomberg*, 9 de junho de 2023, https://www.bloomberg.com/graphics/2023-generative-ai-bias; Victoria Turk, "How AI Reduces the World to Stereotypes", *Rest of World*, 10 de dezembro de 2023, https://restofworld.org/2023/ai-image-stereotypes; Nitasha Tiku, Kevin Schaul e Szu Yu Chen, "This Is How AI Image Generators See the World", *Washington Post*, 1º de novembro de 2023, https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes.

48 Carmen Drahl, "AI Was Asked to Create Images of Black African Docs Treating White Kids. How'd It Go?", *Goats and Soda*,

NPR, 6 de outubro de 2023, https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it.

49 Ezra Klein, "What if Dario Amodei Is Right About A.I.?", *The Ezra Klein Show*, podcast, New York Times Opinion, 12 de abril de 2024, https://www.nytimes.com/column/ezra-klein-podcast.

## CAPÍTULO 5: ESCALA DE AMBIÇÃO

1 Cade Metz, *Criadores de gênios: os inovadores que levaram a IA para o Google, o Facebook e o mundo* (Alta Books, 2022); "Geoffrey Hinton | On Working with Ilya, Choosing Problems, and the Power of Intuition", publicado em 20 de maio de 2024, YouTube, 45 min 45 s, https://youtu.be/n4IQOBka8bc.

2 Entrevista da autora com Geoffrey Hinton, novembro de 2023.

3 Cade Metz, *Criadores de gênios: os inovadores que levaram a IA para o Google, o Facebook e o mundo* (Alta Books, 2022).

4 Will Douglas Heaven, "Rogue superintelligence and merging with machines: Inside the mind of OpenAI's chief scientist", *MIT Technology Review*, 26 de outubro de 2023, https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-and-for-the-future-of-ai/.

5 "NIPS: Oral Session 4 — Ilya Sutskever", publicado em 19 de agosto de 2016, YouTube, 23 min 14 s, https://www.youtube.com/watch?v=-uyXE7dY5H0.

6 Sutskever costuma falar sobre como sua crença na aprendizagem profunda é realmente uma crença. Em setembro de 2023, ele disse: "A criação da OpenAI já foi uma expressão dessa aposta, da ideia de que a aprendizagem profunda consegue fazer isso. Você só precisa acreditar. E, na verdade, eu diria que *muito* do que você sabe, da pesquisa em aprendizagem profunda, pelo menos na última década, talvez um pouco menos agora, tem sido sobre fé." The Open University of Israel, Entrevista com o dr. Ilya Sutskever, cofundador da OpenAI — nos estúdios da Open University — em inglês, publicada em

13 de setembro de 2019, YouTube, 50 min 28 s, https://youtu.be/H1YoNlz2LxA.

7 Khosla Ventures, "What AI Is Making Possible | Ilya Sutskever and Sven Strohband", publicado em 18 de julho de 2023, YouTube, 25 min 26 s, https://youtu.be/xym5f0XY1Sc;Seremot, "Ilya Sutskever: 'Sequence to Sequence Learning with Neural Networks: What a Decade'", publicado em 14 de dezembro de 2024, YouTube, 24 min 36 s, https://youtu.be/1yvBqasHLZs.

8 Entrevista da autora com Pieter Abbeel, outubro de 2023.

9 Seremot, "Ilya Sutskever: 'Sequence to Sequence Learning with Neural Networks: What a Decade'", publicado em 14 de dezembro de 2024, YouTube, 24 min 36 s, https://youtu.be/1yvBqasHLZs.

10 James Incent, "OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'", *The Verge*, 15 de março de 2023, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview.

11 Ilya Sutskever (@ilyasut), "pode ser que as grandes redes neurais de hoje sejam levemente conscientes", Twitter (atual X), 9 de fevereiro de 2022, https://x.com/ilyasut/status/1491554478243258368.

12 Murray Shanahan (@mpshanahan) , "no mesmo sentido em que pode ser que um grande campo de trigo seja levemente macarrão", Twitter (atual X), 10 de fevereiro de 2022, https://x.com/mpshanahan/status/1491715721289678848.

13 Nirit Weiss-Blatt, "What Ilya Sutskever really wants", *AI Panic*, 16 de setembro de 2023, https://aipanic.news/p/what-ilya-sutskever-really-wants.

14 Ilya Sutskever (@ilyasut), "No futuro, quando a robustez dos nossos modelos ultrapassar certo limite, teremos terapias com IA extremamente eficazes e baratas. Isso levará a uma melhoria radical na experiência de vida das pessoas. É uma das aplicações que eu mais aguardo com expectativa", Twitter (atual X), 27 de setembro de 2023, https://x.com/ilyasut/status/1707027536150929689.

15 Uma foto de Sutskever no evento.

16 Ashish Vaswani et al., "Attention is all you need", em *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 6000-6010, dezembro de 2017, https://dl.acm.org/doi/10.5555/3295222.3295349.

17 A tese de doutorado de Sutskever foi focada em redes neurais recorrentes. As RNNs, assim como os Transformers, são projetadas para processar dados sequenciais, o que pode ter aplicações amplas. Por exemplo: uma frase em inglês é uma sequência de palavras, uma imagem é uma sequência de pixels, um vídeo é uma sequência de imagens. Sua tese de doutorado pode ser encontrada em: Ilya Sutskever, "Training recurrent neural networks" (tese de doutorado, Universidade de Toronto, 2013), https://www.cs.utoronto.ca/~ilya/pubs/ilya_sutskever_phd_thesis.pdf.

18 Alec Radford, Karthik Narasimhan, Tim Salimans e Ilya Sutskever, "Improving language understanding by generative pre-training", preprint, *OpenAI*, p. 1-12, 11 de junho de 2018, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. O conjunto de dados original do qual Radford extraiu os mais de 7 mil livros não publicados vem de: Yukun Zhu et al., "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books", em *Proceedings: 2015 IEEE International Conference on Computer Vision* (Institute of Electrical and Electronics Engineers, 2015), p. 19-27, doi.org/10.1109/ICCV.2015.11. Na pesquisa em IA, não é incomum que um grupo reúna um conjunto de dados, disponibilizando-o publicamente, e que outros grupos o reutilizem para os próprios fins.

19 "Generative models", *OpenAI* (blog), 16 de junho de 2016, https://openai.com/index/generative-models.

20 Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg e Dario Amodei, "Deep reinforcement learning from human preferences", em *Proceedings of the 31st International Conference on Neural Information Processing Systems,* p. 4302-4310, dezembro de 2017, dl.acm.org/doi/10.5555/3294996.3295184.

21 "Learning from human preferences", *OpenAI* (blog), 13 de junho de 2017, https://openai.com/index/learning-from-human-preferences.

22 Entrevista da autora com Dario Amodei, agosto de 2019.

23 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei e Ilya Sutskever, "Language models are unsupervised multitask learners", preprint, *OpenAI*, p. 1-24, 14 de fevereiro de 2019, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learne

24 Jared Kaplan et al., "Scaling laws for neural language models", preprint,*arXiv*, 23 de janeiro de 2020, p. 1-30, https://doi.org/10.48550/arXiv.2001.08361.

25 Entrevista da autora com Dario Amodei, agosto de 2019.

26 "Better language models and their implications", *OpenAI* (blog), 14 de fevereiro de 2019, https://openai.com/index/better-language-models.

27 Entrevista da autora com Jack Clark, agosto de 2019.

28 Entrevista da autora com Jack Clark.

29 "Better language models and their implications", *OpenAI* (blog), 14 de fevereiro de 2019, https://openai.com/index/better-language-models.

30 Will Knight, "An AI that writes convincing prose risks mass-producing fake News", *MIT Technology Review*, 14 de fevereiro de 2019, https://www.technologyreview.com/2019/02/14/137426/an-ai-tool-auto-generates-fake-news-bogus-tweets-and-plenty-of-gibberish.

31 Entrevista da autora com Jack Clark, agosto de 2019.

32 Karen Hao, "The messy, secretive reality behind OpenAI's bid to save the world", *MIT Technology Review*, 17 de fevereiro de 2020, https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality.

33 Eu participei dessa reunião da equipe de políticas quando estava trabalhando dentro do escritório em agosto de 2019.

34 Helen Toner, "GPT-2 kickstarted the conversation about publication norms in the AI research Community", *Center for Security and Emerging Technology (CSET)*, 1º de maio de 2020, https://cset.georgetown.edu/article/gpt-2-kickstarted-the-about-norms-in-the-ai-research-community; PAI Staff, "Managing the risks of AI research: six recommendations for responsible publication", *Partnership on AI*, 6 de maio de 2021, https://partnershiponai.org/paper/responsible-publication-recommendations.

35 Entrevista da autora com Dario Amodei, em agosto de 2019.

36 Entrevista da autora com Dario Amodei, em agosto de 2019.

37 Cópias de dois desses documentos.

38 A discussão citada é de um dos documentos mencionados anteriormente.

39 Tom Simonite, "OpenAI said its code was risky. Two grads re-created it anyway", *Wired*, 26 de agosto de 2019, https://www.wired.com/story/dangerous-ai-open-source/

40 A OpenAI não divulgou a quantidade de chips que utilizou em seu artigo original sobre o GPT-3, mas, após uma controvérsia com o Google relatada no capítulo 7, informou aos pesquisadores do Google que foram utilizados 10 mil chips, conforme publicaram no seguinte artigo: David Patterson et al., "Carbon emissions and large neural network training", preprint, *arXiv*, 23 de abril de 2021, p. 6, https://doi.org/10.48550/arXiv.2104.10350.

41 Detalhes sobre os dados de treinamento usados no GPT-2 podem ser encontrados no artigo da OpenAI sobre o modelo: Radford et al., "Language Models Are Unsupervised Multitask Learners".

42 A OpenAI não está sozinha nesse aspecto. Em 2023, Alex Reisner, escritor e programador, confirmaria que empresas como Meta e Bloomberg haviam treinado seus modelos em outro conjunto de dados de livros chamado Books3, que, segundo sua análise, contém mais de 170 mil livros publicados. Em 2024, Reisner também confirmou que as mesmas empresas, inclusive a Anthropic, Nvidia, Apple e outras, estavam treinando seus modelos de forma semelhante

em um conjunto de dados chamado OpenSubtitles, composto dos diálogos de mais de 53 mil filmes e 85 mil episódios de programas de TV. Alex Eisner,"Revealed: The authors whose pirated books are powering generative AI", *The Atlantic*, 19 de agosto de 2023, https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/. Alex Reisner, "There's no longer any doubt that Hollywood writing is powering AI", *The Atlantic*, 18 de novembro de 2024, https://www.theatlantic.com/technology/archive/2024/11/opensubtitles-ai-data-set/680650.

43 Authors Guild v. OpenAI Inc., No. 1:23-cv-08292, *CourtListener* (S.D.N.Y. 6 de maio de 2024) ECF No. 143, Exhibit D, at 2.

44 Para obter detalhes sobre os dados de treinamento utilizados no GPT-3, ver: Tom B. Brown et al., "Language models are few-shot learners", em *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, p. 1877-1901, dezembro de 2020, dl.acm.org/doi/abs/10.5555/3495724.3495883.

45 Entrevista da autora com Ryan Kolln, outubro de 2023.

46 Abeba Birhane, Vinay Prabhu, Sang Han e Vishnu Naresh Boddeti, "On hate scaling laws for data-swamps", preprint, *arXiv*, 28 de junho de 2023, p. 1-27 https://doi.org/10.48550/arXiv.2306.13141.

47 David Thiel, "Identifying and eliminating CSAM in generative ML training data and models", *Stanford Internet Observatory*, p. 1-19, 2023, https://purl.stanford.edu/kh752sm9123.

48 Billy Perrigo, "Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic", *Time*, 18 de janeiro de 2023, https://time.com/6247678/openai-chatgpt-kenya-workers.

49 Karen Hao e Deepa Seetharaman, "Cleaning up ChatGPT takes heavy toll on human Workers", *The Wall Street Journal*, 24 de julho de 2023, https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483; cópia das

instruções de RLHF da OpenAI.

50 Entrevista da autora com Hito Steyerl, em setembro de 2023.

## CAPÍTULO 6: ASCENSÃO

1 Tad Friend, "Sam Altman's Manifest Destiny", *New Yorker*, 3 de outubro de 2016, newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny.

2 "Advice to Entrepreneurs | Sam Altman & Jack Altman", publicado em 1º de agosto de 2019 por Khosla Ventures, YouTube, 30 min 10 s, youtu.be/NAaRhXQCt9o.

3 "Competition Is for Losers with Peter Thiel (How to Start a Startup 2014: 5)", publicado em 22 de março de 2017 por Y Combinator, YouTube, 50 min 27 s, youtu.be/3Fx5Q8xGU8k.

4 "Sam Altman Startup School Video", publicado em 26 de julho de 2017 por Waterloo Engineering, YouTube, 1 h 18 min 19 s, youtu.be/4SlNgM4PjvQ.

5 Tyler Cowen, apresentador, *Conversations with Tyler*, podcast, episódio 61, "Sam Altman on Loving Community, Hating Coworking, and the Hunt for Talent", Mercatus Center Podcasts, 27 de fevereiro de 2019.

6 Entrevista da autora com Geoff Ralston em março de 2024.

7 Cópia do memorando.

8 Stephanie Palazzolo, Erin Woo e Amir Efrati, "How Anthropic Got Inside OpenAI's Head", *The Information*, 12 de dezembro de 2024, https://www.theinformation.com/articles/how-anthropic-got-inside-openais-head.

9 Detalhes das paranoias de Altman e Sutskever e da forma como a empresa intensificou a segurança digital e física foram tirados das recordações e anotações de pessoas que conversaram com Altman ou tinham conhecimento das medidas e de gravações dessas medidas sendo testadas ou discutidas. A ênfase de Altman em segurança também é mencionada no memorando citado anteriormente. Cada detalhe (por exemplo, o foco em ameaças internas, a biometria palmar, as senhas desesperadas) é corroborado por ao menos duas pessoas,

anotações da época, uma gravação ou o memorando.

10 Tom B. Brown, Benjamin Mann, Nick Ryder et al., "Language Models Are Few-Shot Learners", em *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems,* 2020, p. 1877-1901, dl.acm.org/doi/abs/10.5555/3495724.3495883.

11 Além de entrevistas com fontes, a ideia de utilizar modelos de geração de código para acelerar a pesquisa da OpenAI aparece em dois dos memorandos internos da empresa dos quais tenho cópias.

12 Rakesh Kochhar, "Unemployment Rose Higher in Three Months of COVID-19 Than It Did in Two Years of the Great Recession", *Pew Research Center*, 11 de junho de 2020, pewresearch.org/short-reads/2020/06/11/unemployment-rose-higher-in-three-months-of-covid19-than-it-did-in-two-years-of-the-great-recession.

13 Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan et al., "Towards a Human-Like Open-Domain Chatbot", preprint, *arXiv*, 27 de fevereiro de 2020, p. 1-38, doi.org/10.48550/arXiv.2001.09977.

14 Miles Kruppa e Sam Schechner, "How Google Became Cautious of AI and Gave Microsoft an Opening", *The Wall Street Journal*, 7 de março de 2023, wsj.com/articles/google-ai-chatbot-bard-chatgpt-rival-bing-a4c2d2ad.

15 O artigo ganhou um dos prêmios de Melhor Artigo na NeurIPS em 2020. Hsuan-Tien Lin, Maria Florina Balcan, Raia Hadsell e Marc'Aurelio Ranzato, "Announcing the NeurIPS 2020 Award Recipients", *Neural Information Processing Systems Conference*, 8 de dezembro de 2020, neuripsconf.medium.com/announcing-the-neurips-2020-award-recipients-73e4d3101537.

16 *Simple Sabotage Field Manual* (Office of Strategic Services: 1944), cia.gov/static/5c875f3ec660e092cf893f60b4a288df/SimpleSabotage.pdf.

## CAPÍTULO 7: A CIÊNCIA EM CATIVEIRO

1 Cópia do memorando.

2 Karen Hao, Salvador Rodriguez e Deepa Seetharaman, "Mark Zuckerberg Was Early in AI. Now Meta Is Trying to Catch Up", *The Wall Street Journal*, 17 de junho de 2023, wsj.com/articles/mark-zuckerberg-was-early-in-ai-now-meta-is-trying-to-catch-up-94a86284.

3 Jeffrey Ding e Jenny W. Xiao, *Recent Trends in China's Large Language Model Landscape*, Centre for the Governance of AI, 28 de abril de 2023, p. 1-14, cdn.governance.ai/Trends_in_Chinas_LLMs.pdf.

4 Raffaele Huang e Karen Hao, "Baidu Hurries to Ready China's First ChatGPT Equivalent Ahead of Launch", *The Wall Street Journal*, 9 de março de 2023, wsj.com/articles/baidu-scrambles-to-ready-chinas-first-chatgpt-equivalent-ahead-of-launch-bf359ca4.

5 Emma Strubell, Ananya Ganesh e Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, julho de 2019, p. 3645-3650, doi.org/10.18653/v1/P19-1355.

6 David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild et al., "Carbon Emissions and Large Neural Network Training", preprint, *arXiv*, 23 de abril de 2021, doi.org/10.48550/arXiv.2104.10350.

7 Joy Buolamwini e Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", em *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018, p. 77-91, proceedings.mlr.press/v81/buolamwini18a.html.

8 O artigo seguinte: Inioluwa Deborah Raji e Joy Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products", em *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, janeiro de 2019, p. 429-435, doi.org/10.1145/3306618.3314244; a auditoria do governo dos EUA: Patrick Grother, Mei Ngan e Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, NISTIR 8280, National Institute of Standards and Technology, dezembro de 2019, doi.org/10.6028/

NIST.IR.8280.

9 A história completa da pesquisa e do ativismo de Buolamwini é relatada em seu livro de memórias best-seller: Joy Buolamwini, *Unmasking AI: My Mission to Protect What Is Human in a World of Machines* (Random House Trade Paperbacks, 2024); e no documentário da Netflix: *Coded Bias*, dirigido por Shalini Kantayya (2020; Brooklyn, NY: 7th Empire Media), Netflix. Para saber mais sobre os abrangentes impactos de "Gender Shades" e "Actionable Auditing", veja: "Celebrating 5 Years of Gender Shades", Algorithmic Justice League, acessado em 15 de janeiro de 2025, gs.ajl.org/.

10 Karen Hao, "Inside the Fight to Reclaim AI from Big Tech's Control", *MIT Technology Review*, 14 de junho de 2021, technologyreview.com/2021/06/14/1026148/ai-big-tech-timnit-gebru-paper-ethics.

11 Entrevista da autora com Timnit Gebru em agosto de 2023.

12 Alex Hern, "Facebook Translates 'Good Morning' into 'Attack Them,' Leading to Arrest", *The Guardian*, 24 de outubro de 2017, theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest.

13 Safiya Umoja Noble, *Algoritmos da opressão: como o Google fomenta e lucra com o racismo* (Editora Rua do Sabão, 2021).

14 No artigo do GPT-3, na Seção 6.2 "Fairness, Bias, and Representation", são discutidos diversos tipos de vieses encontrados no modelo, seguidos da frase: "Apresentamos esta análise preliminar para compartilhar alguns dos vieses que encontramos, a fim de motivar pesquisas futuras." Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal et al., "Language Models Are Few-Shot Learners", em *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, n. 159, 2020, p. 1877-1901, dl.acm.org/doi/abs/10.5555/3495724.3495883.

15 O relato das experiências de Gebru em torno do artigo "Papagaios Aleatórios" vem principalmente de entrevistas da autora com Gebru, realizadas entre 2020 e 2024,

incluindo uma no dia seguinte a sua demissão, além de uma descrição detalhada em: Tom Simonite, "What Really Happened When Google Ousted Timnit Gebru", *Wired*, 8 de junho de 2021, wired.com/story/google-timnit-gebru-ai-what-really-happened.

16 Diálogo entre Gebru e Emily M. Bender extraído de capturas de tela das conversas, fornecidas por Bender.

17 Cópia do e-mail fornecida por Bender.

18 Simonite, "What Really Happened".

19 Emily M. Bender, Timnit Gebru, Angelina McMillan-Major e Shmargaret Shmitchell [Meg Mitchell], "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", em *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, março de 2021, p. 610-623, doi.org/10.1145/3442188.3445922. Como o Google não permitiu que Meg Mitchell publicasse o artigo pelos motivos detalhados neste capítulo, ela assinou sob o pseudônimo Shmargaret Shmitchell e criou um endereço de e-mail correspondente. Como afiliação, declarou ser "da Aether".

20 Casey Newton, "The Withering Email That Got an Ethical AI Researcher Fired at Google", *Platformer*, 3 de dezembro de 2020, platformer.news/the-withering-email-that-got-an-ethical.

21 Google Walkout for Real Change, "Standing with Dr. Timnit Gebru — #ISupportTimnit #BelieveBlackWomen", *Medium*, 3 de dezembro de 2020, googlewalkout.medium.com/standing-with-dr-timnit-isupporttimnit-believeblackwomen-6dadc300d382.

22 Karen Hao, "We Read the Paper That Forced Timnit Gebru out of Google. Here's What It Says", *MIT Technology Review*, 4 de dezembro de 2020, technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru.

23 Ina Fried, "Scoop: Google CEO Pledges to Investigate Exit of Top AI Ethicist", *Axios*, 9 de dezembro de 2020, axios.com/2020/12/09/sundar-pichai-memo-timnit-gebru-exit.

24 Karen Hao, “Congress Wants Answers from Google About Timnit Gebru’s Firing”, *MIT Technology Review*, 17 de dezembro de 2020, technologyreview.com/2020/12/17/1014994/congress-wants-answers-from-google-about-timnit-gebrus-firing.

25 Ina Fried, “Google Fires Another AI Ethics Leader”, *Axios*, 19 de fevereiro de 2021, axios.com/2021/02/19/google-fires-another-ai-ethics-leader.

26 Sam Shead, “New Google Union ‘Concerned’ After a Senior A.I. Ethics Researcher Is Reportedly Locked Out of Her Account”, *CNBC*, 21 de janeiro de 2021, cnbc.com/2021/01/21/margaret-mitchell-google-investigating-ai-researcher-awu-concerned.html.

27 Sepi Hejazi Moghadam, “Marian Croak’s Vision for Responsible AI at Google”, *The Keyword*, 18 de fevereiro de 2021, blog.google/technology/ai/marian-croak-responsible-ai.

28 Correspondência da autora com porta-voz do Google em novembro de 2024.

29 Mohamed Abdalla e Moustafa Abdalla, “The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity”, em *AIES ’21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, julho de 2021, p 287-297, doi.org/10.1145/3461702.3462563.

30 James Somers, “The Friendship That Made Google Huge”, *New Yorker*, 3 de dezembro de 2018, newyorker.com/magazine/2018/12/10/the-friendship-that-made-google-huge.

31 Simonite, “What Really Happened”.

32 Entrevista da autora com Emma Strubell em novembro de 2023.

33 Correspondência com porta-voz do Google em novembro de 2024.

34 David Patterson, “Good News About the Carbon Footprint of Machine Learning Training”, *Google Research* (blog), 15 de fevereiro de 2022, research.google/blog/good-news-about-the-carbon-footprint-of-machine-learning-training.

35 Correspondência com porta-voz do Google em novembro de 2024.

36 Nitasha Tiku e Gerrit De Vynck, “Google

Shared AI Knowledge with the World — Until ChatGPT Caught Up", *Washington Post*, 4 de maio de 2023, washingtonpost.com/technology/2023/05/04/google-ai-stop-sharing-research.

37 Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong et al., *The Foundation Model Transparency Index,* Stanford Center for Research on Foundation Models, outubro de 2023, crfm.stanford.edu/fmti/October-2023/index.html.

## CAPÍTULO 8: O NASCIMENTO DO COMÉRCIO

1 Cópia do *road map*.

2 Chamado coloquialmente de "Chinchilla paper": Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford et al., "Training Compute-Optimal Large Language Models", preprint, *arXiv*, 29 de março de 2022, p. 1-36, arxiv.org/abs/2203.15556.

3 O primeiro uso deste termo no contexto da IA vem do artigo Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims", preprint, *arXiv*, 20 de abril de 2020, doi.org/10.48550/arXiv.2004.07213.

4 Khlaaf escreveu um artigo que analisa as diferenças entre *red teaming* em IA e segurança. Heidy Khlaaf, "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems", *Trail of Bits*, 7 de março de 2023, p. 1-30, trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

5 Lex Fridman, apresentador, *Lex Fridman Podcast*, podcast, episódio 121, "Eugenia Kuyda: Friendship with an AI Companion", 5 de setembro de 2020, lexfridman.com/podcast.

6 Tom Simonite, "It Began as an AI-Fueled Dungeon Game. It Got Much Darker", *Wired*, 5 de maio de 2021, wired.com/story/ai-fueled-dungeon-game-got-much-darker/.

7 Charles Duhigg, "The Inside Story of Microsoft's Partnership with OpenAI", *New*

*Yorker*, 1° de dezembro de 2023, newyorker.com/magazine/2023/12/11/the-inside-story-of-microsofts-partnership-with-openai.

8 Nat Friedman, "Introducing GitHub Copilot: Your AI Pair Programmer", *GitHub*, 29 de junho de 2021, github.blog/news-insights/product-news/introducing-github-copilot-ai-pair-programmer.

9 OpenAI, "OpenAI Codex", *OpenAI* (blog), 10 de agosto de 2021, openai.com/index/openai-codex.

10 Tiernan Ray, "Microsoft Has Over a Million Paying Github Copilot Users: CEO Nadella", *ZDNet*, 25 de outubro de 2023, zdnet.com/article/microsoft-has-over-a-million-paying-github-copilot-users-ceo-nadella.

11 "Advice to Entrepreneurs | Sam Altman & Jack Altman", publicado em 1º de agosto de 2019 por Khosla Ventures, YouTube, 30 min 10 s, youtu.be/NAaRhXQCt9o.

12 Ellen Huet e Gillian Tan, "Sam Altman Wants to Scan Your Eyeball in Exchange for Cryptocurrency", *Bloomberg*, 29 de junho de 2021, bloomberg.com/news/articles/2021-06-29/sam-altman-s-worldcoin-will-give-free-crypto-for-eyeball-scans.

13 Sarah Holder e Shirin Ghaffary, "Sam Altman–Backed Group Completes Largest US Study on Basic Income", *Bloomberg*, 22 de julho de 2024, bloomberg.com/news/articles/2024-07-22/ubi-study-backed-by-openai-s-sam-altman-bolsters-support-for-basic-income.

14 OpenResearch, "Key Findings: Spending", *OpenResearch* (blog), 21 de julho de 2024, openresearchlab.org/findings/key-findings-spending.

15 Huet e Tan, "Sam Altman Wants to Scan Your Eyeball".

16 Eileen Guo e Adi Renaldi, "Deception, Exploited Workers, and Cash Handouts: How Worldcoin Recruited Its First Half a Million Test Users", *MIT Technology Review*, 6 de abril de 2022, technologyreview.com/2022/04/06/1048981/worldcoin-cryptocurrency-biometrics-web3.

17 Anita Nkonge, "Worldcoin Suspended in Kenya as Thousands Queue for Free Money",

*BBC*, 3 de agosto de 2023, bbc.com/news/world-africa-66383325.

18 Antonio Regalado, "Sam Altman Invested $180 Million into a Company Trying to Delay Death", *MIT Technology Review*, 8 de março de 2023, technologyreview.com/2023/03/08/1069523/sam-altman-investment-180-million-retro-biosciences-longevity-death.

19 Antonio Regalado, "A Startup Is Pitching a Mind-Uploading Service That Is '100 percent Fatal'",*MIT Technology Review*, 13 de março de 2018, technologyreview com/2018/03/13/144721/a-startup-is-pitching-a-mind-uploading-service-that-is-100-percent-fatal.

20 "Office Hours with Sam Altman", publicado em 11 de janeiro de 2017 por Y Combinator, YouTube, 24 min 34 s, youtu.be/45BvnJgwYjk.

21 "StrictlyVC in Conversation with Sam Altman, Part One", publicado em 16 de janeiro de 2023 por Connie Loizos, YouTube, 20 min 32 s, youtu.be/57OU18cogJI.

22 Justine Calma, "Microsoft Just Made a Huge, Far-from-Certain Bet on Nuclear Fusion", *The Verge*, 10 de maio de 2023, theverge.com/2023/5/10/23717332/microsoft-nuclear-fusion-power-plant-helion-purchase-agreement.

23 Informações podem ser encontradas no próprio site: openai.fund.

24 Berber Jin, Tom Dotan e Keach Hagey, "The Opaque Investment Empire Making OpenAI's Sam Altman Rich", *The Wall Street Journal*, 3 de junho de 2024, wsj.com/tech/ai/openai-sam-altman-investments-004fc785.

## CAPÍTULO 9: O CAPITALISMO DO DESASTRE

1 Karen Hao e Deepa Seetharaman, "Cleaning Up ChatGPT Takes Heavy Toll on Human Workers", *The Wall Street Journal*, 24 de julho de 2023, wsj.com/tech/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483.

2 Cópias dos *Statements of Work* (*Declarações de Trabalho*) da OpenAI para o projeto.

3 Entrevista da autora com porta-voz da OpenAI em junho de 2023.

4 Revisão do e-mail.

5 Revisão das respostas.

6 Cópia de dois contratos e revisão dos outros dois.

7 Com base na viagem da autora a Nairóbi para reportagem em maio de 2023.

8 Entrevistas da autora com Mercy Mutemi, advogada que representou quatro trabalhadores quenianos na luta por reformas no trabalho digital no Quênia, em maio de 2023; e Jonathan Beardsley, na época, executivo na empresa de anotação de dados CloudFactory, em maio de 2023.

9 Billy Perrigo, "Inside Facebook's African Sweatshop", *Time*, 14 de fevereiro de 2022, time.com/6147458/facebook-africa-content-moderation-employee-treatment.

10 Correspondência da autora com porta-voz da Sama em novembro de 2024.

11 Caroline Kimeu, "'A Watershed': Meta Ordered to Offer Mental Health Care to Moderators in Kenya", *The Guardian*, 7 de junho de 2023, theguardian.com global-development/2023jun/07/a-watershed-meta-ordered-to-offer-mental-health-care-to-moderators-in-kenya.

12 Com base nos contratos e documentos do projeto, assim como na resposta da Sama ao pedido de comentários da autora para sua reportagem no *Wall Street Journal*: Hao e Seetharaman, "Cleaning Up ChatGPT Takes Heavy Toll".

13 Entrevistas da autora com quatro desses trabalhadores — Mophat Okinyi, Richard Mathenge, Alex Kairu e Bill Mulinya — em 2023.

14 Cópia das instruções que os trabalhadores receberam. Essas categorias correspondem às disponíveis na API de moderação de conteúdo da OpenAI, que podem ser vistas aqui: "Moderation", OpenAI Platform, *OpenAI*, acesso em 17 de outubro de 2024, platform.openai.com/docs/guides/moderation.

15 Entrevistas da autora com Mophat, seu irmão Albert, um de seus amigos e Mutemi, em 2023.

16 Mary L. Gray e Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Harper Business, 2019), p. 1-288; e entrevista da autora com Mary L. Gray em maio de 2019.

17 Florian Alexander Schmidt, "Crowdsourced Production of AI Training Data—How Human Workers Teach Self-Driving Cars How to See", *Working Paper Forschungsförderung*, n. 155, 2019, hdl.handle.net/10419/216075.

18 Entrevistas da autora com Florian Alexander Schmidt, em 2022; e Julian Posada, em 2021.

19 De acordo com o Fundo Monetário Internacional.

20 Schmidt, "Crowdsourced Production of AI".

21 Julian Posada, "The Coloniality of Data Work: Power and Inequality in Outsourced Data Production for Machine Learning" (tese de doutorado, Universidade de Toronto, 2022), p. 1-229, hdl.handle.net/1807/126388.

22 Karen Hao e Andrea Paola Hernández, "How the AI Industry Profits from Catastrophe", *MIT Technology Review*, 20 de abril de 2022, technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels.

23 Entrevistas da autora com Oskarina Veronica Fuentes Anaya, inclusive em sua casa, em 2021.

24 Entrevista da autora com Wilson Pang em dezembro de 2021.

25 Entrevistas da autora com trabalhadores de anotação de dados entre 2021 e 2024 no Quênia, Filipinas, Colômbia, Venezuela (em parceria com Andrea Paola Hernández), norte da África e outros lugares.

26 O relato das práticas comerciais da Scale

baseia-se em entrevistas da autora com cinco funcionários atuais e antigos da empresa, em capturas de tela de documentos internos, análises de instruções fornecidas aos trabalhadores e imersão no Discord do qual fazem parte, além de quase duas dezenas de entrevistas com trabalhadores ao redor do mundo que atuaram na plataforma.

27 Ashlee Vance, "Silicon Valley's Latest Unicorn Is Run by a 22-Year-Old", Bloomberg, em 5 de agosto de 2019, bloomberg.com/news/articles/2019-08-05/scale-ai-is-silicon-valley-s-latest-unicorn.

28 Cópia da planilha com pagamentos de trabalhadores.

29 Correspondência da autora com porta-voz da Scale em novembro de 2024.

30 Correspondência com porta-voz da Scale.

31 Hao e Hernández, "How the AI Industry Profits from Catastrophe".

32 Correspondência com porta-voz da Scale em novembro de 2024.

33 Captura de tela dos pagamentos do trabalhador.

34 Captura de tela da mensagem no canal do Discord dos trabalhadores.

35 Entrevistas da autora com o fundador Mark Sears, em maio de 2023; com o executivo Jonathan Beardsley; com cerca de uma dúzia de trabalhadores da CloudFactory; além de uma visita à sede da CloudFactory em Nairóbi em maio de 2023.

36 Entrevistas da autora com Mophat Okinyi, em maio de 2023; e com Albert Okinyi, em maio e junho de 2023.

37 Segundo a Federação de Empregadores do Quênia, que define "jovens" como pessoas entre 15 e 34 anos.

38 "Continued Rebound, but Storms Cloud the Horizon: Policies to Accelerate the Productive Economy for Inclusive Growth", *Kenya Economic Update*, n. 26, World Bank, 2022, p. 1-54, hdl.handle.net/10986/38386.

39 Correspondência da autora com um porta-voz da Sama em junho de 2023.

40 Entrevistas da autora com Mophat, em maio de 2023; Albert, em maio e junho de 2023; e um amigo de Mophat, em maio de 2023.

41 Cópia do Statement of Work da OpenAI

com a Sama.

42 Cópia das instruções.

43 Mais tarde, pesquisadores da OpenAI escreveram um artigo explicando algumas das práticas que utilizaram para construir o filtro de moderação de conteúdo. A Seção 3.3 aborda como geraram dados sintéticos para o treinamento. O artigo também explica as categorias de severidade. Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler et al., "A Holistic Approach to Undesired Content Detection in the Real World", em *AAAI'23/IAAI'23/EAAI'23: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, n. 1683, 2022, p. 15009-18, dl.acm.org/doi/10.1609/aaai.v37i12.26752.

44 Correspondência com porta-voz da Sama em junho de 2023.

45 Correspondência com porta-voz da Sama.

46 Entrevista com Albert Okinyi, em maio e junho de 2023.

47 Milagros Miceli e Julian Posada, "The Data-Production Dispositif", em *Proceedings of the ACM on Human-Computer Interaction*, v. 6, n. 460, novembro de 2022, p. 1-37, dl.acm.org/doi/10.1145/3555561; James Muldoon e Boxi A. Wu, "Artificial Intelligence in the Colonial Matrix of Power", *Philosophy and Technology*, v. 36, n. 80, dezembro de 2023, doi.org/10.1007/s13347-023-00687-8.

48 Entrevista com Sears em maio de 2023.

49 Acordos da OpenAI baseados em captura de tela de contratos encerrados entre OpenAI e Scale; receita estimada em 2023 de Cory Weinberg, "Fame, Feud and Fortune: Inside Billionaire Alexandr Wang's Relentless Rise in Silicon Valley", *The Information*, 28 de junho de 2024, theinformation.com/articles/fame-feud-and-fortune-inside-billionaire-alexandr-wangs-relentless-rise-in-silicon-valley.

50 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin et al., "Training Language Models to Follow Instructions with Human Feedback", *arXiv*, 4 de março de 2022, p.

1-68, doi.org/10.48550/arXiv.2203.02155.

51 OpenAI, "Aligning Language Models to Follow Instructions", *OpenAI* (blog), 27 de janeiro de 2022, openai.com/index/instruction-following.

52 Baseado em cópias de mais de cem páginas de documentos de RLHF da OpenAI.

53 Documentos de RLHF.

54 Documentos de RLHF.

55 Documentos de RLHF.

56 "John Schulman — Reinforcement Learning from Human Feedback: Progress and Challenges", publicado em 19 de abril de 2023 pela UC Berkeley EECS, YouTube, 1 h 3 min 31 s, youtu.be/hhiLw5Q_UFg.

57 Berber Jin, "The 27-Year-Old Billionaire Whose Army Does AI's Dirty Work", *The Wall Street Journal*, 20 de setembro de 2024, wsj.com/tech/ai/alexandr-wang-scale-ai-d7c6efd7.

58 Alexandr Wang (@alexandr_wang), "empresas começarão a gastar centenas de milhões ou bilhões de dólares em RLHF, assim como em computação para treinar LLMs tipo o ChatGPT... por exemplo, a OpenAI contratou + de 1000 devs para treinar seus modelos de código... uma loucura. Mas em breve as empresas vão gastar centenas de milhões de dólares, ou bilhões, em RLHF, tanto quanto em computação", Twitter (atual X), 1º de fevereiro de 2023, x.com/alexandr_wang/status/1620934510820093952.

59 Correspondência com porta-voz da Scale em novembro de 2024.

60 Baseado em visitas às casas de três trabalhadores da Remotasks e de quatro da Sama em Nairóbi, em maio de 2023, assim como nos endereços de outros dois trabalhadores da Remotasks.

61 Entrevistas da autora com Winnie e sua parceira, Millicent, em maio de 2023.

62 Análise das instruções do Flamingo Generation.

63 Análise das instruções do Crab Generation.

64 Cópia das instruções do Crab Paraphrase.

65 Russell Brandom, "Scale AI's Remotasks Platform Is Dropping Whole Countries Without Explanation", *Rest of World*, 28 de

março de 2024, https://restofworld.org/2024/scale-ai-remotasks-banned-workers/.

66 Jin, "The 27-Year-Old Billionaire".

67 Capturas de tela de designações de grupos e de um anúncio sobre a mudança nos grupos.

68 Cory Weinberg, "Why a $14 Billion Startup Is Now Hiring PhD's to Train AI from Their Living Rooms", *The Information*, 25 de junho de 2024, theinformation.com/articles/why-a-14-billion-startup-is-now-hiring-phds-to-train-ai-from-their-living-rooms.

69 Hilary Kimuyu, "Online Gig Site Remotasks Exits Kenya", *Business Daily*, 13 de março de 2024, businessdailyafrica.com/bd/corporate/technology/online-gig-site-remotasks-exits-kenya-4555340.

## CAPÍTULO 10: DEUSES E DEMÔNIOS

1 Andrew Van Dam, "What Percent Are You?", *Economics Blog*, *The Wall Street Journal*, 2 de março de 2016, graphics.wsj.com/what-percent/.

2 Tyler Cowen, apresentador, *Conversations with Tyler*, podcast, episódio 61, "Sam Altman on Loving Community, Hating Coworking, and the Hunt for Talent", Mercatus Center Podcasts, 27 de fevereiro de 2019.

3 Émile P. Torres, "The Acronym Behind Our Wildest AI Dreams and Nightmares", *Truthdig*, 15 de junho de 2023, truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares/.

4 William MacAskill, "Replaceability, Career Choice, and Making a Difference", *Ethical Theory and Moral Practice*, v. 17, 2013, p. 269-283, doi.org/10.1007/S10677-013-9433-4.

5 "What Is Effective Altruism?", *Effective Altruism Forum*, acesso em 8 de outubro de 2024, effectivealtruism.org/articles/introduction-to-effective-altruism.

6 Will MacAskill, "What Are the Most Important Moral Problems of Our Time?", *TED Talk*, abril de 2018, 11 min 45 s, ted.com/talks/will_macaskill_what_are_the_most_important_moral

7 "About Us", *Open Philanthropy*, acesso em 17 de outubro de 2024, openphilanthropy.org/about-us.

8 Holden Karnofsky, "The Open Philanthropy Project Is Now an Independent Organization", *Open Philanthropy*, 12 de junho de 2017, openphilanthropy.org/research/the-open-philanthropy-project-is-now-an-independent-organization.

9 David Yaffe-Bellany, "A Crypto Emperor's Vision: No Pants, His Rules", *New York Times*, 14 de maio de 2022, nytimes.com/2022/05/14/business/sam-bankman-fried-ftx-crypto.html.

10 Rebecca Ackermann, "Inside Effective Altruism, Where the Far Future Counts a Lot More Than the Present", *MIT Technology Review*, 17 de outubro de 2022, technologyreview.com/2022/10/17/1060967/effective-altruism-growth.

11 " Announcing the Future Fund", *FTX Future Fund*, arquivado em 27 de novembro de 2022, no web.archive.org/web/20221127183608/https://ftxfuturefund.org/announcing-the-future-fund/.

12 "An Overview of the AI Safety Funding Situation", *Effective Altruism Forum*, acesso em 8 de outubro de 2024, forum.effectivealtruism.org/posts/XdhwXppfqrpPL2YDX/an-overview-of-the-ai-safety-funding-situation; correspondência do autor com o porta-voz da Open Philanthropy em novembro de 2024.

13 Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff e Mona Wang, "Building the Epistemic Community of AI Safety", preprint, *SSRN*, 1º de dezembro de 2023, p. 1-14, ssrn.com/abstract=4641526; "What Is Effective Altruism?", *Effective Altruism Forum*.

14 A maior parte dessas definições foram retiradas dos blogs *LessWrong* e *Effective Altruism Forum*; por exemplo, "AI Timelines", *LessWrong*, acesso em 17 de outubro de 2024, lesswrong.com/tag/ai-timelines; "Global Catastrophic Risk", *Effective Altruism Forum*, acesso em 27 de novembro de 2024, forum.effectivealtruism.org/topics/global-catastrophic-risk.

15 Charlotte Alter, "Effective Altruism Promises to Do Good Better. These Women Say It Has a Toxic Culture of Sexual

Harassment and Abuse", *Time*, 3 de fevereiro de 2023, time.com/6252617/effective-altruism-sexual-harassment; Kelsey Piper, "Why Effective Altruism Struggles on Sexual Misconduct", *Vox*, 16 de fevereiro de 2023, vox.com/future-perfect/2023/2/15/23601143/effective-altruism-sexual-harassment-misconduct.

16 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal et al., "Learning Transferable Visual Models from Natural Language Supervision", preprint, *arXiv*, 26 de fevereiro de 2021, p. 1-48, doi.org/10.48550/arXiv.2103.00020.

17 OpenAI, "DALL·E: Creating Images from Text", *OpenAI* (blog), 5 de janeiro de 2021, openai.com/index/dall-e.

18 Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan e Surya Ganguli, "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics", *nos anais do ICML 2015: Proceedings of the 32nd International Conference on Machine Learning*, v. 37, julho de 2015, p. 2256-65, dl.acm.org/doi/10.5555/3045118.3045358.

19 Jonathan Ho, Ajay Jain e Pieter Abbeel, "Denoising Diffusion Probabilistic Models", em *NIPS 2020: Proceedings of the 34th International Conference on Neural Information Processing Systems*, n. 574, dezembro de 2020, p. 6840-51, dl.acm.org/doi/abs/10.5555/3495724.3496298; Anil Ananthaswamy, "The Physics Principle That Inspired Modern AI Art", *Quanta Magazine*, 5 de janeiro de 2023, quantamagazine.org/the-physics-principle-that-inspired-modern-ai-art-20230105.

20 "DALL·E 2", *OpenAI*, acesso em 17 de setembro de 2024, openai.com/index/dall-e-2.

21 Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models", em *Proceedings of the 39th International Conference on Machine Learning*, 2022, p. 16784-804, proceedings.mlr.press/v162/nichol22a.html.

22 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser e Björn

Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", em *2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 10674-85, doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

23 Entrevista da autora com Björn Ommer em março de 2024.

24 Fraser Kelton e Nabeel Hyatt, apresentadores, *Hallway Chat*, podcast, "Launch Stories of ChatGPT", 2 de dezembro de 2023, hallwaychat.co/launch-stories-of-chatgpt.

25 Hayden Field, "Microsoft Engineer Warns Company's AI Tool Creates Violent, Sexual Images, Ignores Copyrights", *CNBC*, 6 de março de 2024, cnbc.com/2024/03/06/microsoft-ai-engineer-says-copilot-designer-creates-disturbing-images.html.

26 Kelton e Hyatt, *Hallway Chat*.

27 Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson e Nico Grant, "How Tech Giants Cut Corners to Harvest Data for A.I.", *New York Times*, 6 de abril de 2024, nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

28 Davey Alba e Emily Chang, "YouTube Says OpenAI Training Sora with Its Videos Would Break Rules", *Bloomberg*, 4 de abril de 2024, bloomberg.com/news/articles/2024-04-04/youtube-says-openai-training-sora-with-its-videos-would-break-the-rules.

29 OpenAI, "Introducing Whisper", *OpenAI* (blog), 21 de setembro de 2022, openai.com/index/whisper.

30 "GPT-4 Contributions", *OpenAI*, acesso em 13 de outubro de 2024, openai.com/contributions/gpt-4.

31 Bill Gates, apresentador, *Unconfuse Me with Bill Gates*, podcast, episódio 2, "Sal Khan", *Gates Notes*, 10 de agosto de 2023, gatesnotes.com/podcast.

32 Bill Gates, "The Age of AI Has Begun", *GatesNotes*, 21 de março de 2023, gatesnotes.com/The-Age-of-AI-Has-Begun.

33 Repetido publicamente por Bill Gates, incluindo em "The Age of AI Has Begun".

34 Kelton e Hyatt, *Hallway Chat*.

35 Will Hurd, "Should 4 People Be Able to Control the Equivalent of a Nuke?", *Politico*, 30 de janeiro de 2024, politico.com/news/magazine/2024/01/30/will-hurd-ai-regulation-00136941.

36 "Sam Altman Startup School Video", postado em 26 de julho de 2017, por Waterloo Engineering, YouTube, 1 h 18 min 19 s, youtu.be/4SlNgM4PjvQ.

37 Bilawal Sidhu, apresentador, *The TED AI Show*, podcast, "What Really Went Down at OpenAI and the Future of Regulation w/ Helen Toner", 28 de maio de 2024, ted.com/talks/the_ted_ai_show_what_really_went_down_at_opena

38 Rebecca Heilweil, "Why Silicon Valley Is Fertile Ground for Obscure Religious Beliefs", *Vox*, 30 de junho de 2022, vox.com/recode/2022/6/30/23188222/silicon-valley-blake-lemoine-chatbot-eliza-robot.

39 Nitasha Tiku, "The Google Engineer Who Thinks the Company's AI Has Come to Life", *Washington Post*, 11 de junho de 2022, washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

40 Tom Hartsfield, "Koko the Impostor: Ape Sign Language Was a Bunch of Babbling Nonsense", *Big Think*, 11 de maio de 2022, bigthink.com/life/ape-sign-language.

41 Entrevista da autora com Geoff Hinton.

## CAPÍTULO 11: ÁPICE

1 A foto em questão, outubro de 2022.

2 United States v. Samuel Bankman-Fried, No. 1:22-cr-00673, CourtListener (S.D.N.Y., 15/3/2024), ECF nº 410, pp. 12–13. O trecho relevante diz: "Do final de 2021 até o primeiro trimestre de 2022, Bankman-Fried direcionou bilhões de dólares em gastos, que utilizaram dinheiro de clientes da FTX. Essas despesas incluíam […] a Anthropic PBC (uma empresa de inteligência artificial)."

3 Zack Abrams, "FTX Offloads Remaining Anthropic Shares as Bankruptcy Cost Surpasses $500 Million", *The Block*, 1º de junho de 2024, theblock.co/post/298010/ftx-remaining-anthropic-shares-as-bankruptcy-cost-700-million.

4 Will Douglas Heaven, "The Inside Story of How ChatGPT Was Built from the People Who Made It", *MIT Technology Review*, 3 de março de 2023, technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai.

5 "StrictlyVC in Conversation with Sam Altman, Part Two", publicado em 17 de janeiro de 2023 por Connie Loizos, YouTube, 38 min 58s, youtu.be/bjkD1Om4uw.

6 Erin Woo e Stephanie Palazzolo, "OpenAI Overhauls Content Moderation Efforts as Elections Loom", *The Information*, 18 de dezembro de 2023, theinformation.com/articles/openai-overhauls-content-moderation-efforts-as-elections-loom.

7 "Behind the Scenes Scaling ChatGPT — Evan Morikawa at LeadDev West Coast 2023", publicado em 26 de outubro de 2023, por LeadDev, YouTube, 27 min 12 s, youtu.be/PeKMEXUrlq4.

8 OpenAI, "Using GPT-4 for Content Moderation", *OpenAI* (blog), 15 de agosto de 2023, openai.com/index/using-gpt-4-for-content-moderation.

9 Nico Grant e Cade Metz, "A New Chat Bot Is a 'Code Red' for Google's Search Business", *New York Times*, 21 de dezembro de 2022, nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html.

10 Cópia do memorando.

11 O relato sobre o clima na Microsoft baseia-se em entrevistas da autora com dez funcionários e executivos atuais e antigos da empresa, além de cópias de vários e-mails enviados por executivos aos funcionários.

12 Cópia do e-mail que faz referência à nova estratégia.

13 Cada um dos e-mails da Microsoft citados é baseado em cópias desses e-mails.

14 O crescimento nos pedidos de inferência baseia-se em cópias dos e-mails acima mencionados, bem como em capturas de tela de um painel interno.

15 Correspondência da autora com um porta-voz da Microsoft, em novembro de 2024, que forneceu essa citação a partir de uma transcrição da reunião.

16 Woo e Palazzolo, "OpenAI Overhauls Content Moderation".

17 Woo e Palazzolo, "OpenAI Overhauls Content Moderation".

18 Cópia do documento.

19 Dylan Patel e Afzal Ahmad, "The Inference Cost of Search Disruption — Large Language Model Cost Analysis", *SemiAnalysis*, 9 de fevereiro de 2023, semianalysis.com/p/the-inference-cost-of-search-disruption.

20 Jon Victor e Aaron Holmes, "OpenAI Dropped Work on New 'Arrakis' AI Model in Rare Setback", *The Information*, 17 de outubro de 2023, theinformation.com/articles/openai-dropped-work-on-new-arrakis-ai-model-in-rare-setback.

21 Tom Dotan e Deepa Seetharaman, "The Awkward Partnership Leading the AI Boom", *The Wall Street Journal*, 13 de junho de 2023, wsj.com/articles/microsoft-and-openai-forge-awkward-partnership-as-techs-new-power-couple-3092de51.

22 Karen Weise e Cade Metz, "How Microsoft's Satya Nadella Became Tech's Steely Eyed A.I. Gambler", *New York Times*, 14 de julho de 2026, nytimes.com/2024/07/14/technology/microsoft-ai-satya-nadella.html.

23 Anissa Gardizy e Amir Efrati, "Microsoft and OpenAI Plot $100 Billion Stargate AI Supercomputer", *The Information*, 29 de março de 2024, theinformation.com/articles/microsoft-and-openai-plot-100-billion-stargate-ai-supercomputer; Anissa Gardizy, Aaron Holmes e Amir Efrati, "OpenAI Leaders Say Microsoft Isn't Moving Fast Enough to Supply Servers", *The Information*, 8 de outubro de 2024, theinformation.com/articles/openai-eases-away-from-microsoft-data-centers.

## CAPÍTULO 12: TERRA SAQUEADA

1 Baseado na viagem de reportagem da autora a Santiago e ao deserto do Atacama, 2024.

2 Entrevista da autora com Sonia Ramos, uma ativista atacamenha, em junho de 2024; o corte de línguas também é mencionado na introdução de um dicionário de kunza, uma língua atacamenha extinta: Julio Vilte Vilte, *Kunza: Lengua del Pueblo Lickan Antai o*

*Atacameño* (Codelco Chile, 2004), p. 11.

3 "Chile — Country Commercial Guide: Mining", *International Trade Administration*, 7 de dezembro de 2023, trade.gov/country-commercial-guides/chile-mining.

4 Samo Burja, "Chile Is a Politically Disunited Resource Exporter", *Bismarck Brief*, 19 de junho de 2024, brief.bismarckanalysis.com/p/chile-is-a-politically-disunited.

5 Naomi Klein, *A doutrina do choque: a ascensão do capitalismo do desastre* (Nova Fronteira, 2008).

6 Naomi Klein, *A doutrina do choque: a ascensão do capitalismo do desastre* (Nova Fronteira, 2008).

7 Milton Friedman, "A Friedman Doctrine — The Social Responsibility of Business Is to Increase Its Profits", *The New York Times*, 13 de setembro de 1970, timesmachine.nytimes.com/timesmachine/1970/09/13/223535702.html?pageNumber=379.

8 Naomi Klein, *A doutrina do choque: a ascensão do capitalismo do desastre* (Nova Fronteira, 2008).

9 James Doubek, "The U.S. Set the Stage for a Coup in Chile. It Had Unintended Consequences at Home", NPR, 10 de setembro de 2023, npr.org/2023/09/10/1193755188/chile-coup-50-years-pinochet-kissinger-human-rights-allende; o relatório original do Senado que detalha os gastos pesados e a campanha de influência da CIA no Chile antes do golpe: Covert Action in Chile 1963-1973, *Covert Action in Chile 1963– 1973, Staff Report of the Select Committee to Study Governmental Operations with Respect to Intelligence Activities* (US Senate: 1975), intelligence.senate.gov/sites/default/files/94chile.pdf.

10 Daniel Matamala, "The Complicated Legacy of the 'Chicago Boys' in Chile", *Promarket*, 12 de setembro de 2021, promarket.org/2021/09/12/Chicago-boys-chile-friedman-neoliberalism.

11 "Income Share of the Richest 1%", *Our World in Data*, acesso em 14 de outubro de 2024, ourworldindata.org/grapher/income-share-top-1-before-tax-wid?tab=chart&country=CHL.

12 Gobierno de Chile, "International InvestChile Forum: 100 Companies from 28 Countries Will Meet in the Country", 16 de maio de 2024, gob.cl/en/news/international-investchile-forum-100-companies-from-28-countries-will-meet-in-the-country/.

13 Entrevista da autora com Martín Tironi Rodó em junho de 2024.

14 Entrevistas da autora com Alan Howard, analista de nuvem e data centers da consultoria Omdia, entre agosto e setembro de 2023.

15 De fato era, até a autora visitar o data center que treinava modelos da OpenAI no Arizona, em setembro de 2023.

16 Entrevista da autora com Mél Hogan em agosto de 2023.

17 Bianca Bosker, "Why Everything Is Getting Louder", *The Atlantic*, 15 de novembro de 2019, theatlantic.com/magazine/archive/2019/11/the-end-of-silence/598366.

18 Rich Miller, "The Gigawatt Data Center Campus Is Coming", *Data Center Frontier*, 29 de abril de 2024, datacenterfrontier.com/hyperscale/article/55021675/the-gigawatt-data-center-campus-is-coming.

19 Entrevistas da autora com Hogan em agosto de 2023; e com um investidor de data center em março de 2024.

20 Goldman Sachs, "AI Is Poised to Drive 160% Increase in Data Center Power Demand", 14 de maio de 2024, goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand.

21 Uma instalação de 150 megawatts pode consumir até 150 MWh em uma hora, ou 1.314.000 MWh por ano. Segundo a Administração de Informações sobre Energia dos Estados Unidos, uma casa norte-americana consumiu, em média, 10.791 kWh em 2022; 1.314.000 MWh divididos por 10.791 kWh equivalem a 121.768.

22 Uma instalação de 1.000 megawatts pode consumir até 8.760.000 MWh/ano; uma de 2.000 megawatts, o dobro. A cidade de São Francisco consumiu 5.120.586 MWh em 2022; assim, 8.760.000/5.120.586 = 1,7. O dobro disso é 3,4. "Electricity Consumption by County", Califórnia Energy Commission, acesso em 17 de outubro de 2024.

23 Goldman Sachs, "AI Is Poised to Drive 160% Increase".

24 Evan Halper, "A Utility Promised to Stop Burning Coal. Then Google and Meta Came to Town", *The Washington Post*, 12 de outubro de /10/2024, washingtonpost.com/business/2024/10/08/google-meta-omaha-data-centers/; C. Mandler, "Three Mile Island Nuclear Plant Will Reopen to Power Microsoft Data Centers", NPR, 20 de setembro de 2024, npr.org/2024/09/20/nx-s1-5120581/three-mile-island-nuclear-power-plant-microsoft-ai.

25 Goldman Sachs, "AI Is Poised to Drive 160% Increase"; Ian King, "AI Computing on Pace to Consume More Energy Than India, Arm Says", *Bloomberg*, 17 de abril de 2024, news.bloomberglaw.com/artificial-intelligence/ai-computing-on-pace-to-consume-more-energy-than-india-arm-says.

26 Esta é uma afirmação que Altman fez muitas vezes, incluindo em: Sam Altman, "The Intelligence Age", *Sam Altman* (blog), 23 de setembro de 2024, ia.samaltman.com.

27 Entrevista da autora com Sasha Luccioni em agosto de 2023.

28 Climate Change AI detalha essas tecnologias em vários relatórios em seu site, climatechange.ai, incluindo David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran et al., "Tackling Climate Change with Machine Learning", *ACM Computing Surveys* (CSUR), v. 55, n. 2, fevereiro de 2022, p. 1-96, doi.org/10.1145/3485128.

29 Alexandra Sasha Luccioni, Yacine Jernite e Emma Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?", em *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, junho de 2024, p. 85-99, doi.org/10.1145/3630106.3658542.

30 Esses números são baseados na Tabela 2 do artigo mencionado e na estimativa da EPA, anterior a janeiro de 2024, de que uma recarga de smartphone consumia 0,012 kWh de energia.

31 Transcrição da reunião.

32 Entrevista com Alan Howard agosto de 2023.

33 James Glanz, "How the Internet Kept

Humming During 2 Hurricanes", *New York Times*, 18 de setembro de 2017, nytimes.com/2017/09/18/us/harvey-irma-internet.html.

34 Pengfei Li, Jianyi Yang, Mohammad A. Islam e Shaolei Ren, "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models", preprint, *arXiv*, 29 de outubro de 2023, p. 1, doi.org/10.48550/arXiv.2304.03271.

35 Md Abu Bakar Siddik, Arman Shehabi e Landon Marston, "The Environmental Footprint of Data Centers in the United States", *Environmental Research Letters*, v. 16, n. 6, junho de 2021, p. 064017, doi.org/10.1088/1748-9326/abfba1.

36 Entrevistas da autora com seis comunidades diferentes que enfrentavam expansões de data centers no Arizona, no Novo México, na Virgínia, duas no Chile e Uruguai, entre 2023 e 2024, assim como entrevistas com três funcionários e executivos da Microsoft, incluindo Noelle Walsh, vice-presidente corporativa de operações e inovação em nuvem, que supervisiona todas as expansões de data centers da empresa, sobre as práticas da companhia sob sua perspectiva, entre 2023 e 2024.

37 Entrevista da autora com Roger Yackel, um morador da Virgínia que lidera protestos contra a expansão de data centers, em março de 2024.

38 Cópia do e-mail.

39 Entrevista da autora com Greg Brockman em agosto de 2019.

40 Os codinomes, números e localização das Fases 1, 2 e 3 são retirados de um documento interno da OpenAI. As localizações e os custos das Fases 4 e 5 vêm de: Anissa Gardizy e Amir Efrati, "Microsoft and OpenAI Plot $100 Billion Stargate AI Supercomputer", *The Information*, 29 de março de 2024, theinformation.com/articles/microsoft-and-openai-plot-100-billion-stargate-ai-supercomputer.

41 Matt O'Brien e Hannah Fingerhut, "Artificial Intelligence Technology Behind ChatGPT Was Built in Iowa — with a Lot of Water", *AP*, 9 de setembro de 2023, https://apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-

f551fde98083d17a7e8d904f8be822c4.

42 Correspondência da autora com porta-voz da Microsoft em novembro de 2024.

43 Entrevistas da autora com Barbara Chappell, diretora dos serviços de água da cidade de Goodyear, em outubro de 2023; dois moradores da comunidade, em setembro de 2023; as três fontes da Microsoft mencionadas anteriormente, entre 2023 e 2024; e cópias das atas do conselho municipal de Goodyear e outros documentos e correspondências governamentais obtidos por meio de pedidos de registros públicos, entre 2023 e 2024. Essas entrevistas e reportagens adicionais resultaram na seguinte matéria: Karen Hao, "AI Is Taking Water from the Desert", *The Atlantic*, 1º de março de 2024, theatlantic.com/technology/archive/2024/03/ai-water-climate-microsoft/677602/.

44 A estimativa de 5 mil megawatts para o Stargate vem de Gardizy e Efrati, "Microsoft and OpenAI Plot $100 Billion Stargate"; e, de acordo com o Gabinete do Prefeito de Nova York para Justiça Climática e Ambiental, a cidade usou em média cerca de 5,5 mil megawatts de energia em 2022: "Systems", NYC Mayor's Office of Climate and Environmental Justice, acesso em 17 de outubro de 2024, climate.cityofnewyork.us/subtopics/systems.

45 Berber Jin, Tom Dotan e Keach Hagey, "The Opaque Investment Empire Making OpenAI's Sam Altman Rich", *The Wall Street Journal*, 3 de junho de 2024, wsj.com/tech/ai/openai-sam-altman-investments-004fc785.

46 Segundo o West Des Moines Water Works, conforme citado por: O'Brien e Fingerhut, "Artificial Intelligence Technology Behind ChatGPT".

47 Correspondência com porta-voz da Microsoft em novembro de 2024.

48 A. Park Williams, Benjamin I. Cook e Jason E. Smerdon, "Rapid Intensification of the Emerging Southwestern North American Megadrought in 2020-2021", *Nature Climate Change*, v. 12, n. 3, março de 2022, p. 232–34, doi.org/10.1038/s41558-022-01290-z.

49 Isso se refere a uma condição chamada "deadpooling", conforme explicado em: Christopher Flavelle e Mira Rojanasakul, "As the Colorado River Shrinks, Washington

Prepares to Spread the Pain", *New York Times*, 27 de janeiro de 2023, nytimes.com/2023/01/27/climate/colorado-river-biden-cuts.html.

50 Kira Caspers, "645 People Died Due to Heat in Metro Phoenix in 2023. Here's What Is Changing This Year", *AZ Central*, 15 de março de 2024, azcentral.com/story/news/local/phoenix/2024/03/15/heat-deaths-maricopa-county/72980594007/.

51 Entrevista da autora com Tom Buschatzke em outubro de 2023.

52 Kevin Lee, Adi Gangidi e Mathew Oldham, "Building Meta's GenAI Infrastructure", *Engineering at Meta*, 12 de março de 2024, https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/.

53 Entrevista com Sonia Ramos em junho de 2024.

54 "Tres muertos y treinta heridos en explosión de una mina en Chuquicamata", *El Mercurio*, 6 de setembro de 1957.

55 "The Battle for Chile's Critical Minerals", publicado em 22 de julho de 2022 pela Sky News, YouTube, 13 min 54 s, youtu.be/oywE0mQnWI0.

56 Entrevista da autora com Cristina Dorador, uma cientista chilena que estuda os ecossistemas do Deserto do Atacama, em junho de 2024.

57 "The Battle for Chile's Critical Minerals", Sky News; entrevista com Dorador.

58 Visitas e entrevistas da autora com três líderes atacamenhos, incluindo Sonia Ramos e Sergio Cubillos, em junho de 2024.

59 Visitas e entrevistas com os três líderes atacamenhos; e visita a uma clínica de saúde patrocinada pela indústria, em junho de 2024.

60 Entrevistas da autora com Dorador em junho de 2024; e com a SQM, empresa chilena de mineração e maior produtora de lítio do mundo; além de uma visita técnica às minas de lítio da SQM no Atacama, em junho de 2024.

61 Govind Bhutada, "This Chart Shows Which Countries Produce the Most Lithium", *Fórum Econômico Mundial* (blog), 5 de maio de 2023, weforum.org/stories/2023/01/chart-

countries-produce-lithium-world/.

62 Entrevistas da autora com Dorador em junho de 2024; com SQM, junho de 2024; e com a arquiteta e pesquisadora Marina Otero Verzier, em maio de 2024, que apresenta a seguinte palestra sobre a relação entre a extração de lítio, o desenvolvimento de data centers, a história colonial do Chile e futuros tecnológicos globais: "Marina Otero Verzier — Data Mourning", publicado em 1º de março de 2023 por Columbia GSAPP, YouTube, 1 h 30 min, youtu.be/vbFPaNBNB-M.

63 Visita e entrevista com Cubillos, líder de Peine, em junho de 2024.

64 Entrevista com SQM em junho de 2024.

65 Entrevista com Dorador em junho de 2024.

66 Paul R. La Monica, "Move Over, Nvidia. Copper Is Getting a Big AI Boost Too", *Barron's*, 22 de maio de 2024, barrons.com/articles/copper-price-ai-microsoft-utilities-c99058b7.

67 "Artificial Intelligence, Art and Indigeneity", acesso em 2 de outubro de 2024, aei.art.br/aiai/en/the-research.

68 Visitas e entrevistas com os três líderes atacamenhos em junho de 2024.

69 Dylan Patel e Myron Xie, "Microsoft Infrastructure — AI & CPU Custom Silicon Maia 100, Athena, Cobalt 100", *SemiAnalysis*, 15 de novembro de 2023, https://newsletter.semianalysis.com/p/microsoft-infrastructure-ai-and-cpu.

70 Alphabet, "2024 Q3 Earnings Call", *Alphabet Investor Relations*, 29 de outubro de 2024, abc.xyz/2024-q3-earnings-call.

71 Meta, "Meta Reports Third Quarter 2024 Results", *Meta Investor Relations*, 30 de outubro de 2024, investor.atmeta.com/investor-news/press-release-details/2024/Meta-Reports-Third-Quarter-2024-Results/default.aspx.

72 Com base na visita da autora a Quilicura em junho de 2024.

73 Correspondências da autora com a porta-voz do Google Chile entre junho e novembro de 2024.

74 Entrevistas da autora com Alexandra Arancibia em junho de 2024.

75 Entrevistas da autora com Arancibia em

junho de 2024; Rodrigo Vallejos, em junho de 2024; Lorena Antiman, outra ativista ambiental em Quilicura, em junho de 2024; Miguel Mora, um professor de Quilicura que estuda seus pântanos, e Felipe Gonzalez, que chefia a Unidade de Gestão Ambiental de Quilicura, em junho de 2024.

76 Entrevistas com Arancibia, Vallejos e Antiman.

77 Subsecretaría de Telecomunicaciones, "Gobierno de Chile escoge ruta mediante Nueva Zelanda y hasta Australia para implementar el Cable Transoceánico", *Subsecretaria de Telecomunicações*, 27 de julho de 2020, subtel.gob.cl/gobierno-de-chile-escoge-ruta-mediante-nueva-zelanda-y-hasta-australia-para-implementar-el-cable-transoceanico/.

78 Google, "Announcing Humboldt, the First Cable Route Between South America and Asia-Pacific", *Google Cloud* (blog), 11 de janeiro de 2024, https://cloud.google.com/blog/products/infrastructure/announcing-humboldt-the-first-cable-route-between-south-america-and-asia-pacific.

79 Josefa Silva González, "A más de 20 años de Miño: La estancada lucha contra el asbestos", *La Voz de Maipú*, 18 de fevereiro de 2022, https://lavozdemaipu.cl/la-estancada-lucha-contra-el-asbesto/.

80 Entrevistas com Arancibi, Vallejos e Gonzalez; e entrevista da autora com Tania Rodriguez em junho de 2024.

81 O relatório de impacto ambiental da Google ao SEA afirmava que o data center poderia usar 169 litros de água potável por segundo, ou 5.329.584.000 litros por ano. Segundo a autoridade de abastecimento de água de Cerillos, o município consumiu 5.097.946 litros durante todo o ano de 2019, quando a Google buscava se instalar; 5.329.584.000 litros por ano divididos por 5.097.946 litros anuais equivalem a 1.045 litros.

82 "Persistent Drought Is Drying Out Chile's Drinking Water", *Reuters*, 20 de março de 2024, reuters.com/world/americas/persistent-drought-is-drying-out-chiles-drinking-water-2024-03-20/#:~:text=The%20historic%20drought%20has%20impacted,of%20water%20supply%20by%202040..

83 O relato do ativismo do MOSACAT contra

a Google baseia-se em entrevistas da autora com Rodriguez e outros oito membros do MOSACAT, em junho de 2024. Detalhes adicionais vêm da cobertura da mídia chilena, principalmente de Alberto Arellano, Lucas Cifuentes e Cristóbal Ríos, "Las zonas oscuras de la evaluación ambiental que autorizó 'a ciegas' el megaproyecto de Google en Cerrillos", *Ciper*, 25 de maio de 2020, ciperchile.cl/2020/05/25/las-zonas-oscuras-de-la-evaluacion-ambiental-que-autorizo-a-ciegas-el-megaproyecto-de-google-en-cerrillos.

84 Todos os detalhes sobre as operações da Antel baseiam-se na visita da autora a um data center da Antel e em uma entrevista com seu gerente, Javier Echeverria, em junho de 2024.

85 Entrevista da autora com Marcos Umpiérrez, professor da Universidad de la República, no Uruguai, e seus colegas, em junho de 2024.

86 Baseado em visita ao Parque de las Ciencias, em junho de 2024.

87 Entrevista da autora com Marcelo Fozati, agrônomo e agricultor uruguaio que lidera uma organização de defesa de agricultores e cultivos locais, e com Daniel Pena, pesquisador uruguaio que estuda o extrativismo ambiental de multinacionais no país, em junho de 2024; além de "Uruguay: Drought Losses Estimated at USD 1.200 million, Minister Says", *MercoPress*, 2 de fevereiro de 2023, en.mercopress.com/2023/02/02/uruguay-drought-losses-estimated-at-usd-1.200-million-minister-says; e Guillermo Garat, "My City Has Run Out of Fresh Water. Will Your City Be Next?", *New York Times*, 19 de junho de 2023, nytimes.com/2023/07/19/opinion/drinking-water-montevideo.html.

88 Entrevistas da autora com três residentes e ativistas da água no Uruguai: Fabiana, em junho de 2024; Noelia Lagos, em junho de 2024; e Carmen Sosa, em junho de 2024.

89 "Google's and Microsoft's Profits Soar as Pandemic Benefits Big Tech", *New York Times*, 18 de outubro de 2021, nytimes.com/live/2021/04/27/business/stock-market-today.

90 Entrevista com Fabiana.

91 Grace Livingstone, "'It's Pillage': Thirsty

Uruguayans Decry Google's Plan to Exploit Water Supply", *Guardian*, 11 de julho 2023, theguardian.com/world/2023/jul/11/uruguay-drought-water-google-data-center.

92 Entrevista com Fozati e Pena.

93 "Fertilizer Use Per Capita, 1961 to 2019", *Our World in Data*, acesso em 17 de outubro de 2024, ourworldindata.org/grapher/fertilizer-per-capita?tab=table.

94 Os detalhes sobre a pesquisa e o ativismo de Pena baseiam-se em entrevistas da autora com Pena, entre maio e junho de 2024, incluindo um dia passado em seu caminhão viajando por algumas das áreas mais pobres de Montevidéu e de sua periferia.

95 Entrevistas com Pena, em junho de 2024; e com Sosa, em junho de 2024.

96 Livingstone, "'It's Pillage'".

97 Livingstone, "'It's Pillage'".

98 Correspondência da autora com a porta-voz do Google Chile em novembro de 2024.

99 Dan Swinhoe, "Microsoft Files Plans for Chilean Data Center Region", *Data Center Dynamics*, 24 de janeiro de 2022, datacenterdynamics.com/en/news/microsoft-files-plans-for-chilean-data-center-region/.

100 Matamala, "The Complicated Legacy of the 'Chicago Boys'".

101 Rodrigo Vallejos Calderón, "Los costos de estar conectados: Datacenters y el consumo hídrico", *Bits 23*, 2022, p. 28-33, revistasdex.uchile.cl/index.php/bits/issue/view/1049.

102 Entrevistas da autora com Marina Otero Verzier em maio de 2024; e Serena Dambrosio e Nicolás Díaz Bejarano, em junho de 2024.

103 Fotos dos modelos.

104 Entrevistas com Martín Tironi e Aisén Etcheverry, chefe do Ministério da Ciência.

## CAPÍTULO 13: OS DOIS PROFETAS

1 "Watch: OpenAI CEO Sam Altman Testifies Before Senate Judiciary Committee", *PBS News*, 16 de maio de 2023, pbs.org/newshour/politics/watch-live-openai-ceo-sam-altman-testifies-before-senate-judiciary-committee.

2 Gary Marcus, "OpenAI's Sam Altman Is

Becoming One of the Most Powerful People on Earth. We Should Be Very Afraid", *Guardian*, 3 de agosto de 2024, theguardian.com/technology/article/2024/aug/03/open-ai-sam-altman-chatgpt-gary-marcus-taming-silicon-valley.

3 Hasan Chowdhury, "Insiders Say Sam Altman's AI World Tour Was a Success", *Business Insider*, 24 de junho de 2023, businessinsider.com/sam-altman-world-tour-ai-chatgpt-openai-2023-6.

4 Cecilia Kang, "How Sam Altman Stormed Washington to Set the A.I. Agenda", *New York Times*, 7 de junho de 2023, nytimes.com/2023/06/07/technology/sam-altman-ai-regulations.html.

5 Kang, "How Sam Altman Stormed Washington".

6 Entrevistas da autora com Karla Ortiz, em dezembro de 2023 e abril de 2024; e com Rachel Meinerding e Nicole Hendrix Herman, as cofundadoras e colíderes da Concept Art Association, em abril de 2024.

7 Entrevista com Meinerding e Hendrix Herman; CVL Economics, *Future Unscripted: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs*, 2024, p. 1-58,animationguild.org/wp-content/uploads/2024/01/Future-Unscripted-The-Impact-of-Generative-Artificial-Intelligence-on-Entertainment-Industry-Jobs-pages-1.pdf.

8 Kang, "How Sam Altman Stormed Washington".

9 Karen Hao, "The New AI Panic", *The Atlantic*, 11 de outubro de 2023, theatlantic.com/technology/archive/2023/10/technology-exports-ai-programs-regulations-china/675605.

10 Alex W. Palmer, "'An Act of War': Inside America's Silicon Blockade Against China", *New York Times*, 12 de julho de 2023, nytimes.com/2023/07/12/magazine/semiconductor-chips-us-china.html.

11 Jane Lee, "Exclusive: Nvidia Offers New Advanced Chip for China That Meets U.S. Export Controls", *Reuters*, 7 de novembro de 2022, reuters.com/technology/exclusive-nvidia-offers-new-advanced-chip-china-that-meets-us-export-controls-2022-11-08.

12 Fanny Potkin e Yelin Mo, "Chinese Chip

Equipment Makers Grab Market Share as US Tightens Curbs", *Reuters*, 18 de outubro de 2023, reuters.com/technology/chinese-chip-equipment-makers-grab-market-share-us-tightens-curbs-2023-10-18/.

13 Khari Johnson, "Meta's Open Source Llama Upsets the AI Horse Race", *Wired*, 26 de julho de 2023, wired.com/story/metas-open-source-llama-upsets-the-ai-horse-race/.

14 Tony Peng, "What Llama 3 Means to China, ERNIE Bot Hits 200 Million Users, and China Trails US in AI Models", *Recode China AI*, 22 de abril de 2024, recodechinaai.substack.com/p/what-llama-3-means-to-china-ernie.

15 Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety", preprint, *arXiv*, 7 de novembro de 2023, p. 1-51, doi.org/10.48550/arXiv.2307.03718.

16 Entrevistas da autora com Sara Hooker, em outubro de 2024; Deborah Raji, em agosto de 2024; Sarah Myers West, codiretora do AI Now, em outubro de 2024; e outros especialistas em políticas de IA, entre 2023 e 2024.

17 Sara Hooker, "On the Limitations of Compute Thresholds as a Governance Strategy", preprint, *arXiv*, 30 de julho de 2024, p. 1-54, doi.org/10.48550/arXiv.2407.05694.

18 Entrevistas da autora com Myers West, em setembro de 2023; Amba Kak, a outra codiretora do AI Now, em outubro de 2023; Emily Weinstein, em setembro de 2023; Raji, em outubro de 2023; e dois outros especialistas em políticas de IA, em novembro de 2023.

19 Entrevistas com Weinstein.

20 Hao, "The New AI Panic".

21 "Watch: OpenAI CEO Sam Altman Testifies."

22 Entrevista com Myers West em setembro de 2023.

23 Departamento de Comércio dos Estados Unidos, "NTIA Solicits Comments on Open-Weight AI Models", press release, 21 de fevereiro de 2024, commerce.gov/news/press-releases/2024/02/ntia-solicits-

comments-open-weight-ai-models.

24 "Mozilla's Response to the National Telecommunications and Information Administration's Request for Comments on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights", *Mozilla Foundation* (blog), março de 2024, blog.mozilla.org/netpolicy/files/2024/03/Mozilla-RfC-Submission-Dual-Use-Foundation-Models-With-Widely-Available-Model-Weights.pdf.

25 Trechos desta seção apareceram de outra forma em Hao, "The New AI Panic".

26 Cameron F. Kerry, Joshua P. Meltzer e Matt Sheehan, "Can Democracies Cooperate with China on AI Research?", *Brookings*, 9 de janeiro de 2023, brookings.edu/articles/can-democracies-cooperate-with-china-on-ai-research/.

27 Matt Sheehan, "Who Benefits from American AI Research in China?", *Macro Polo*, 21 de outubro de 2019, macropolo.org/china-ai-research-resnet.

28 O relato de como a ordem executiva foi elaborada baseia-se em entrevistas da autora com Alondra Nelson, ex-diretora do OSTP, em outubro de 2023; Suresh Venkatasubramanian, ex-diretor adjunto do OSTP, em outubro de 2023; e duas outras pessoas da área de políticas públicas, em novembro de 2023.

29 Trechos desta seção apareceram de outra forma em Karen Hao e Matteo Wong, "The White House Is Preparing for an AI-Dominated Future", *The Atlantic*, 30 de outubro de 2023, theatlantic.com/technology/archive/2023/10/biden-white-house-ai-executive-order/675837.

30 Khari Johnson, "Why Silicon Valley Is Trying So Hard to Kill This AI Bill in Califórnia", *CalMatters*, 12 de agosto de 2024, calmatters.org/economy/technology/2024/08/ai-regulation-showdown.

31 Entrevista com Hooker em outubro de 2024.

32 Gabby Miller, "US Senate AI 'Insight Forum' Tracker", *Tech Policy Press*, 9 de dezembro de 2023, techpolicy.press/us-senate-ai-insight-forum-tracker.

33 Entrevista da autora com Raji em outubro de 2023; Inioluwa Deborah Raji, "AI's Present

Matters More Than Its Imagined Future", *The Atlantic*, 4 de outubro de 2023, theatlantic.com/technology/archive/2023/10/ai-chuck-schumer-forum-legislation/675540.

34 Cat Zakrzewski, "Meet the Woman Who Transformed Sam Altman into the Avatar of AI", *Washington Post*, 9 de janeiro de 2024, washingtonpost.com/technology/2024/01/09/openai-anna-makanju-ai-regulation.

35 Sam Altman (@sama), "farei uma viagem entre maio e junho para falar com usuários e desenvolvedores da openai (e com os interessados em ia, no geral). por favor, venham ficar de bobeira, compartilhar pedidos de funcionalidades e outros feedbacks! mais detalhes aqui: https://openai.com/form/openai-tour-2023 ou email p/ oai23tour@openai.com", Twitter (atual X), 29 de março de 2023, x.com/sama/status/1641181668206858240.

36 "GPT-4 contributions", *OpenAI*, acessado em 13 de outubro de 2024, openai.com/contributions/gpt-4.

37 "GPT-4", *OpenAI*, 14 de março de 2023, openai.com/index/gpt-4-research.

38 Sam Altman (@sama), "o GPT-4 foi realmente um trabalho de equipe, árduo, e de toda a empresa. mas, no geral, a liderança e visão técnica de Jakub Pachocki antes do treinamento foi memorável, e não estaríamos aqui sem ela", Twitter (atual X), 14 de março de 2023, x.com/sama/status/1635700851619819520.

39 OpenAI, "OpenAI and Journalism", *OpenAI* (blog), 8 de janeiro de 2024, openai.com/index/openai-and-journalism.

40 Dan Milmo, "'Impossible' to Create AI Tools like ChatGPT Without Copyrighted Material, OpenAI Says", *Guardian*, 8 de janeiro de 2024, theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai.

41 OpenAI, "Our Approach to AI Safety", *OpenAI* (blog), 5 de abril de 2023, openai.com/index/our-approach-to-ai-safety.

42 OpenAI, "Introducing Superalignment", *OpenAI* (blog), 5 de julho de 2023, openai.com/index/introducing-superalignment.

43 Sam Altman (@sama), "eu estava esperando

que o filme do Oppenheimer inspirasse uma geração de crianças a se tornarem físicas, mas deixou a desejar nesse aspecto", publicou Altman no Twitter. "vamos fazer esse filme acontecer! (acho que *A rede social* conseguiu inspirar os fundadores de startups.)", Twitter (atual X), 22 de julho de 2023, x.com/sama/status/1682809958734131200.

44 Elizabeth Weil, "Sam Altman Is the Oppenheimer of Our Age", *New York*, 25 de setembro de 2023, nymag.com/intelligencer/article/sam-altman-artificial-intelligence-openai-profile.html.

45 Cade Metz, "The ChatGPT King Isn't Worried, but He Knows You Might Be", *New York Times*, 31 de março de 2023, nytimes.com/2023/03/31/technology/sam-altman-open-ai-chatgpt.html.

46 Tyler Cowen, apresentador, *Conversations with Tyler*, podcast, episódio 61, "Sam Altman on Loving Community, Hating Coworking, and the Hunt for Talent", Mercatus Center Podcasts, 27 de fevereiro de 2019.

47 Eliezer Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut It All Down", *Time*, 29 de março de 2023, time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough.

48 Musk v. Altman, No. 4:24-cv-04722, CourtListener (N.D. Cal. 14 de novembro de 2024) ECF No. 32, Evidência 18.

49 Julia Black, "Elon Musk Had Twins Last Year with One of His Top Executives", *Business Insider*, 6 de julho de 2022, businessinsider.com/elon-musk-shivon-zilis-secret-twins-neuralink-tesla.

50 "Elon Musk Wanted an OpenAI For-Profit", *OpenAI* (blog), 13 de dezembro de 2024, openai.com/index/elon-musk-wanted-an-openai-for-profit/#summer-2017-we-and-elon-agreed-that-a-for-profit-was-the-next-step-for-openai-to-advance-the-mission.

51 Sam Altman, "Quora", Sam Altman (blog), 21 de abril de 2017, blog.samaltman.com/quora.

52 Helen Toner, "Leaning into EA Disillusionment", *Effective Altruism Forum*, 22 de julho de 2022, forum.effectivealtruism.org/posts/MjTB4MvtedbLjgyja/leaning-into-ea-

disillusionment.

53 Cade Metz, Tripp Mickle e Mike Isaac, "Before Altman's Ouster, OpenAI's Board Was Divided and Feuding", *New York Times*, 21 de novembro de 2023, nytimes.com/2023/11/21/technology/openai-altman-board-fight.html.

54 Kevin Roose, "OpenAI Insiders Warn of a 'Reckless' Race for Dominance", *New York Times*, 4 de junho de 2024, nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html.

55 Dan Primack, "Sam Altman Owns OpenAI's Venture Capital Fund", *Axios*, 15 de fevereiro de 2024, axios.com/2024/02/15/sam-altman-openai-startup-fund.

## CAPÍTULO 14: LIVRAMENTO

1 Elizabeth Weil, "Sam Altman Is the Oppenheimer of Our Age", *New York*, 25 de setembro de 2023, nymag.com/intelligencer/article/sam-altman-artificial-intelligence-openai-profile.html.

2 Cópia do e-mail, que Annie postou on-line: Annie Altman (@anniealtman108), "Menos de 24 hrs depois da @NYMag publicar o primeiro reconhecimento público e oficial da minha existência relação x.com/bullishdumping/bullishdumping/status/1753869400719958519", Twitter (atual X), 3 de fevereiro de 2024, x.com/anniealtman108/status/1753881201482629258.

3 Entrevistas da autora e visita a Annie Altman entre março e novembro de 2024.

4 Correspondência da autora com Connie Gibstine em outubro de 2024.

5 Altman v. Altman, No. 4:25-cv-00017, CourtListener (E.D. Mo. 6 de janeiro de 2025) ECF No. 1; Sam Altman (@sama), "Minha irmã entrou com um processo contra mim. Aqui está uma declaração minha, de minha mãe e de meus irmãos", Twitter (atual X), 7 de janeiro de 2025, x.com/sama/status/1876780763653263770.

6 Correspondência para verificação de fatos com a Burroughs, em outubro de 2024; entrevista da autora com James Roble, em julho de 2024.

7 Cópia dos registros médicos de Annie em

Tufts.

8 Cada um dos diagnósticos de Annie é corroborado por cópias de um dos seguintes documentos: seus registros médicos da infância; registros médicos da Tufts University; notas de terapia da Tufts University; exames e laudos de imagem da vida adulta, inclusive um ultrassom e uma ressonância magnética; uma avaliação de obstetrícia e ginecologia; e registros de fisioterapia. Detalhes do impacto desses diagnósticos em sua mobilidade e qualidade de vida também são comprovados por seus registros de terapia em Tufts; registros de fisioterapia da vida adulta; e fotos, como de sua bota ortopédica e de lençóis encharcados de suor.

9 "Jerry Altman Obituary", *St. Louis Post-Dispatch*, 27 de maio de 2018, legacy.com/us/obituaries/stltoday/name/jerry-altman-obituary?id=1683283.

10 Cópia dos registros médicos da infância de Annie; cópia das notas de terapia da Tufts University.

11 Trevor Noah, apresentador, *What Now? with Trevor Noah*, temporada 1, episódio 5, "Sam Altman Speaks Out About What Happened at OpenAI", Spotify Podcasts, 7 de dezembro de 2023, open.spotify.com/show/122imavATqSE7eCyXIcqZL.

12 Cópia da notificação por e-mail do fundo de aposentadoria e do extrato do mesmo fundo com o saldo.

13 Cópia do e-mail.

14 Cópia das notas da terapeuta de Annie em Los Angeles.

15 Cópia do e-mail da notificação bancária.

16 Captura de tela do e-mail de ativação do site SeekingArrangement.

17 Cópias de diversos e-mails e mensagens trocados entre Annie e sua família.

18 Cópias de diversos e-mails e mensagens.

19 Cópia da troca de mensagens.

20 Cópia do e-mail.

21 Cópia dos e-mails de ativação dos sites Etsy e Patreon.

22 Diversas capturas de tela.

23 Stories antigos do Instagram de Annie com capturas de tela de suas avaliações no Apple

Podcasts e marcando a equipe de suporte a respeito do desaparecimento delas.

24 Captura de tela de um Instagram no qual o número de curtidas é maior do que o número de visualizações; e duas capturas de tela do mesmo vídeo no YouTube, em que a captura com o carimbo de tempo posterior exibe menor número de visualizações.

25 Entrevistas da autora com Olivia Snow, pesquisadora da UCLA com ênfase em trabalho sexual, tecnologia e políticas públicas, em maio de 2024; Val Elefante, pesquisadora e uma das fundadoras da Reliabl, empresa feminista de mídia social, em setembro de 2024; e um ex-cientista de dados do Facebook, em outubro de 2024.

26 Entrevista da autora com Neily Messerschmidt, em novembro de 2024.

27 Cópia das notas de terapia de Annie no Havaí.

28 Notas de terapia de Annie.

29 Entrevista com Messerschmidt em novembro de 2024.

30 Entrevistas da autora com uma terapeuta, em junho e agosto de 2024, que também fez referência ao livro best-seller: Bessel van der Kolk, *O corpo guarda as marcas: cérebro, mente e corpo na cura do trauma* (Editora Sextante, 2020).

31 Annie Altman (@anniealtman108), "Sofri abuso sexual, físico, emocional, verbal, financeiro e tecnológico por parte de meus irmãos biológicos, principalmente Sam Altman e, em menor grau, Jack Altman. (2/3)", Twitter (atual X), 13 de novembro de 2021, x.com/anniealtman108/status/1459696444802142213.

32 Annie Altman (@anniealtman108), "Sam e Jack, sei que vocês se lembram do torá, minha parte era sobre Moisés perdoando seus irmãos — *perdoe-os, Pai, porque eles não sabem o que fazem*. Abuso sexual, físico, emocional, verbal, financeiro e tecnológico. Jamais esquecido.", Twitter (atual X), 10 de setembro de 2022, x.com/anniealtman108/status/1568689744951005185.

33 Captura de tela do histórico de renda do OnlyFans de Annie.

34 Cópia da troca de e-mails.

35 Cópias do testamento e do trust de Jerry

Altman.

36 Cópia da correspondência por e-mail entre a advogada de Annie e a advogada de Gibstine.

37 Altman, CourtListener, ECF No. 1.

38 Cópia do diagnóstico de Annie.

39 Entrevistas da autora com a terapeuta mencionada anteriormente e com Blaise Aguirre, professor assistente de psiquiatria na Harvard Medical School, em outubro de 2024, ambos com experiência no trabalho com diversos pacientes diagnosticados com o transtorno. Nenhum deles revisou o caso de Annie, apenas comentou sobre a condição de forma geral.

40 A pesquisa mais ampla sobre o transtorno de personalidade borderline é um estudo longitudinal em andamento de 24 anos chamado *McLean Study of Adult Development*, conduzido por Mary C. Zanarini, que acompanhou mais de 360 pessoas diagnosticadas com o transtorno. Entre as principais descobertas do estudo: o transtorno tem bom prognóstico sintomático, e medicações psicotrópicas não são curativas. O estudo publica regularmente novos artigos, incluindo: Mary C. Zanarini, Frances R. Frankenburg, Isabel V. Glass e Garrett M. Fitzmaurice, "The 24-Year Course of Symptomatic Disorders in Patients with Borderline Personality Disorder and Personality-Disordered Comparison Subjects: Description and Prediction of Recovery From BPD", *The Journal of Clinical Psychiatry,* v. 85, 2024, doi.org/10.4088/JCP.24m15370.

## CAPÍTULO 15: O GAMBITO

1 O relato sobre a infância e a juventude de Murati vem principalmente de Charles Duhigg, "The Inside Story of Microsoft's Partnership with OpenAI", *New Yorker*, 1º de dezembro de 2023, newyorker.com/magazine/2023/12/11/the-inside-story-of-microsofts-partnership-with-openai; e da participação de Murati no podcast de Kevin Scott: Kevin Scott, apresentador, *Behind the Tech with Kevin Scott*, "Mira Murati, Chief Technology Officer, OpenAI", Microsoft, 11 de julho de 2023, microsoft.com/en-us/behind-the-tech/mira-murati-chief-technology-officer-openai.

2 Christopher Jarvis, "The Rise and Fall of Albania's Pyramid Schemes", *Finance & Development*, International Monetary Fund, março de 2000, imf.org/external/pubs/ft/fandd/2000/03/jarvis.htm.

3 Duhigg, "The Inside Story of Microsoft's Partnership with OpenAI".

4 Exceto por indicação em contrário, o relato sobre o período que antecedeu a crise do conselho e os bastidores da própria crise, neste e no próximo capítulo, baseia-se em entrevistas da autora com oito pessoas diretamente envolvidas ou próximas às pessoas diretamente envolvidas nos eventos descritos; em suas anotações da mesma época; e em capturas de tela de mensagens no Slack, e-mails e outras evidências, inclusive o áudio da reunião com todos os funcionários de 17 de novembro de 2023, após o conselho ter destituído Altman.

5 Mike Isaac, Tripp Mickle e Cade Metz, "Key OpenAI Executive Played a Pivotal Role in Sam Altman's Ouster", *New York Times*, 7 de março de 2024, nytimes.com/2024/03/07/technology/openai-executives-role-in-sam-altman-ouster.html.

6 Isaac et al., "Key OpenAI Executive Played a Pivotal Role".

7 Cade Metz, Tripp Mickle e Mike Isaac, "Before Altman's Ouster, OpenAI's Board Was Divided and Feuding", *New York Times*, 21 de novembro de 2023, nytimes.com/2023/11/21/technology/openai-altman-board-fight.html.

## CAPÍTULO 16: MISTÉRIO

1 Deepa Seetharaman, Keach Hagey, Berber Jin e Kate Linebaugh, "Sam Altman's Knack for Dodging Bullets — with a Little Help from Bigshot Friends", *The Wall Street Journal*, 24 de dezembro de 2023, wsj.com/tech/ai/sam-altman-openai-protected-by-silicon-valley-friends-f3efcf68.

2 Natasha Mascarenhas, "Behind OpenAI Meltdown, Valley Heavyweight Reid Hoffman Calmed Microsoft Nerves", *The Information*, 17 de janeiro de 2024, theinformation.com/articles/behind-openai-meltdown-valley-heavyweight-reid-hoffman-calmed-microsoft-nerves.

3 Mike Isaac, Tripp Mickle e Cade Metz, "Key

OpenAI Executive Played a Pivotal Role n Sam Altman's Ouster", *New York Times*, 7 de março de 2024, nytimes.com/2024/03/07/technology/openai-executives-role-in-sam-altman-ouster.html.

4 Elon Musk (@elonmusk), "Estou muito preocupado. Ilya tem um bom senso moral e não busca poder. Ele não tomaria uma atitude tão drástica a menos que sentisse que era absolutamente necessário.", Twitter (atual X), 19 de novembro de 2023, x.com/elonmusk/status/1726376406785925566.

5 Elon Musk (@elonmusk), Twitter (atual X), 20 de novembro de 2023, x.com/elonmusk/status/1726542015087927487.

6 Elon Musk (@elonmusk), "Esta carta sobre a OpenAI acabou de ser enviada a mim. Essas parecem ser preocupações que valem ser investigadas." https://gist.github.com/Xe/32d7bc436e401f3323ae77e7e242f858", Twitter (atual X), 21 de novembro de 2023, x.com/elonmusk/status/1727096607752282485.

7 "Xe/openai-message-to-board.md", *GitHub Gist*, arquivado em 21 de novembro de 2023 em web.archive.org/web/20231121225252/https://gist.github.com/Xe/32d7bc436e401f3323ae77e7e242f858.

8 Karen Hao, "The Messy, Secretive Reality Behind OpenAI's Bid to Save the World", *MIT Technology Review*, 17 de fevereiro de 2020, technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-greg-messy-secretive-reality; Karen Hao e Charlie Warzel, "Inside the Chaos at OpenAI", *The Atlantic*, 19 de novembro de 2023, theatlantic.com/technology/archive/2023/11/sam-altman-open-ai-chatgpt-chaos/676050.

9 Todas as citações de e-mails vêm das capturas de tela fornecidas pela pessoa.

10 Anna Tong, Jeffrey Dastin e Krystal Hu, "OpenAI Researchers Warned Board of AI Breakthrough Ahead of CEO Ouster, Sources Say", *Reuters*, 23 de novembro de 2023, reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by-letter-board-about-ai-breakthrough-2023-11-22.

11 Jon Victor e Amir Efrati, "OpenAI Made an AI Breakthrough Before Altman Firing, Stoking Excitement and Concern", *The*

*Information*, 22 de novembro de 2023, theinformation.com/articles/openai-made-an-ai-breakthrough-before-altman-firing-stoking-excitement-and-concern.

12 "Ilya Sutskever: 'Sequence to Sequence Learning with Neural Networks: What a Decade'", publicado em 14 de dezembro de 2024 por seremot, YouTube, 24 min 36 s, youtu.be/1yvBqasHLZs.

13 Anna Tong e Katie Paul, "Exclusive: OpenAI Working on New Reasoning Technology Under Code Name 'Strawberry'", *Reuters*, 15 de julho de 2024, reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12/.

14 Trechos desta seção apareceram de outra forma como Karen Hao, "Why Won't OpenAI Say What the Q* Algorithm Is?", *The Atlantic*, 28 de novembro de 2023, theatlantic.com/technology/archive/2023/11/openai-sam-altman-q-algorithm-breakthrough-project/676163.

15 OpenAI, "Creating Video from Text", *OpenAI* (blog), openai.com/index/sora.

16 Kate Conger e Lauren Hirsch, "The Board Chair Squaring Up to Elon Musk in the Feud Over Twitter", *New York Times*, 4 de outubro de 2022, nytimes.com/2022/10/04/technology/twitter-board-elon-musk.html.

17 "OpenAI Chair's AI Startup Sierra Gets $4.5 Bln Valuation in Latest Funding Round", *Reuters*, 28 de outubro de 2024, reuters.com/technology/artificial-intelligence/openai-chairs-ai-startup-sierra-gets-45-bln-valuation-latest-funding-round-2024-10-28/#:~:text=Oct%2028%20(Reuters)%20%2D%20Sierra,the%20company%20said%20on%20Monday..

18 OpenAI, "Review Completed & Altman, Brockman to Continue to Lead OpenAI", *OpenAI* (blog), 8 de março de 2024, openai.com/index/review-completed-altman-brockman-to-continue-to-lead-openai.

19 OpenAI, "Review Completed".

20 Helen Toner divulgou a declaração em uma captura de tela no X: Helen Toner (@hlntnr), "Uma declaração de Helen Toner e Tasha McCauley:", Twitter (atual X), 8 de março de 2024, x.com/hlntnr/status/1766269137628590185.

## CAPÍTULO 17: ACERTO DE CONTAS

1 Erin Woo e Stephanie Palazzolo, "OpenAI Researchers, Including Ally of Sutskever, Fired for Alleged Leaking", *The Information*, 11 de abril de 2024, theinformation.com/articles/openai-researchers-including-ally-of-sutskever-fired-for-alleged-leaking.

2 Edward Ludlow e Ashlee Vance, "Altman Sought Billions for Chip Venture Before OpenAI Ouster", *Bloomberg*, 19 de novembro de 2023, bloomberg.com/news/articles/2023-11-19/altman-sought-billions-for-ai-chip-venture-before-openai-ouster.

3 Keach Hagey e Asa Fitch, "Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI", *The Wall Street Journal*, 8 de fevereiro de 2024, wsj.com/tech/ai/sam-altman-seeks-trillions-of-dollars-to-reshape-business-of-chips-and-ai-89ab3db0.

4 Lex Fridman, apresentador, *Lex Fridman Podcast*, episódio 419, "Sam Altman: OpenAI, GPT-5, Sora, Board Saga, Elon Musk, Ilya, Power & AGI", 18 de março de 2024, lexfridman.com/podcast.

5 Entrevista da autora com Alexis Conneau em janeiro de 2025.

6 Cópia de memorando interno da OpenAI.

7 OpenAI, *Preparedness Framework (Beta)*, OpenAI, 18 de dezembro de 2023, p. 1-27, cdn.openai.com/openai-preparedness-framework-beta.pdf.

8 OpenAI, *Preparedness Framework (Beta)*.

9 OpenAI, "Hello GPT-4o", *OpenAI* (blog), 13 de maio de 2024, openai.com/index/hello-gpt-4o.

10 Informação confirmada de forma independente por duas pessoas.

11 "Trump's Thirsty VP Contenders Crash Trial & ChatGPT's Flirty AI Update | The Daily Show", publicado em 15 de maio de 2024 pelo The Daily Show, YouTube, 9 min 57 s, youtu.be/eFkUOi_9140.

12 Gravação de áudio da reunião, 15 de maio de 2024.

13 Gravação de áudio da reunião, 15 de maio de 2024.

14 Gravação de áudio de Altman expressando seu arrependimento aos funcionários.

15 Sarah Krouse, Deepa Seetharaman e Joe Flint, "Behind the Scenes of Scarlett Johansson's Battle with OpenAI", *The Wall Street Journal*, 23 de maio de 2024, wsj.com/tech/ai/scarlett-johansson-openai-sam-altman-voice-fight-7f81a1aa.

16 Fridman, "Sam Altman".

17 "Sam Altman & Brad Lightcap: Which Companies Will Be Steamrolled by OpenAI?", publicado em 15 de abril de 2024 por 20VC com Harry Stebbings, YouTube, 53 min 6 s, youtu.be/G8T1O81W96Y.

18 Julia Black, "The Besties' Revenge: How the 'All-In' Podcast Captured Silicon Valley", *The Information*, 15 de dezembro de 2023, theinformation.com/articles/the-besties-revenge-how-the-all-in-podcast-ate-silicon-valley.

20 "In Conversation with Sam Altman", publicado em 10 de maio de 2024 por All-In Podcast, YouTube, 1 h 43 min 2 s, youtu.be/nSM0xd8xHUM.

21 Sam Altman (@sama), "tento não pensar muito nos concorrentes, mas não consigo parar de pensar na diferença estética entre a openai e o google", Twitter (atual X), 16 de maio de 2024, x.com/sama/status/1791183356274921568.

22 Deepa Seetharaman, "SEC Investigating Whether OpenAI Investors Were Misled", *The Wall Street Journal*, 28 de fevereiro de 2024, wsj.com/tech/sec-investigating-whether-openai-investors-were-misled-9d90b411.

23 Karen Weise e Cade Metz, "How Microsoft's Satya Nadella Became Tech's Steely Eyed A.I. Gambler", *New York Times*, 14 de setembro de 2024, nytimes.com/2024/07/14/technology/microsoft-ai-satya-nadella.html.

24 Baseado nas recordações e caracterizações de três pessoas que trabalharam com ele.

25 Rob Copeland e Parmy Olson, "Artificial Intelligence Will Define Google's Future. For Now, It's a Management Challenge", *The Wall Street Journal*, 26 de janeiro de 2021, wsj.com/tech/ai/artificial-intelligence-will-define-googles-future-for-now-its-a-management-challenge-11611676945; Giles Turner e Mark Bergen, "Google DeepMind Co-Founder Placed on Leave From AI Lab", *Bloomberg*, 21 de agosto de 2019,

bloomberg.com/news/articles/2019-08-21/google-deepmind-co-founder-placed-on-eave-from-ai-lab.

26 Jordan Novet, "Microsoft Says OpenAI Is Now a Competitor in AI and Search", *CNBC*, 31 de julho de 2024, cnbc.com/2024/07/31/microsoft-says-openai-is-now-a-competitor-in-ai-and-search.html; Alex Heath, "Microsoft Now Lists OpenAI as a Competitor", *The Verge*, 2 de agosto de 2024, theverge.com/2024/8/2/24212370/microsoft-now-lists-openai-as-a-competitor.

27 Ellen Huet, apresentadora, *Foundering: The OpenAI Story*, podcast, temporada 5, episódio 1, "The Most Silicon Valley Man Alive", Bloomberg Podcasts, 6 de junho de 2024, bloomberg.com/news/audio/2024-06-05/foundering-openai-e1-most-silicon-valley-man-alive-podcast.

28 "Sam Altman Talks GPT-4o and Predicts the Future of AI", publicado em 14 de maio de 2024 pelo Logan Bartlett Show, YouTube, 46 min 14 s, youtu.be/fMtbrKhXMWc.

29 OpenAI, "Ilya Sutskever to Leave OpenAI, Jakub Pachocki Announced as Chief Scientist", *OpenAI* (blog), 14 de maio de 2024, openai.com/index/jakub-pachocki-announced-as-chief-scientist/.

30 Ilya Sutskever (@ilyasut), "Após quase uma década, tomei a decisão de deixar a OpenAI. A trajetória da empresa foi nada menos que milagrosa, e estou confiante de que a OpenAI construirá uma IAG que seja ao mesmo tempo segura e benéfica sob a liderança de @sama, @gdb, @miramurati e, agora, sob a excelente liderança de pesquisa de @merettm. Foi uma honra e um privilégio termos trabalhado juntos, e sentirei muita falta de todos. Até logo, e obrigado por tudo. Estou animado com o que vem a seguir — um projeto que é muito significativo para mim pessoalmente e sobre o qual compartilharei detalhes no momento oportuno", Twitter (atual X), 14 de maio de 2024, x.com/ilyasut/status/1790517455628198322.

31 Captura de tela do anúncio no Slack, em 14 de maio de 2024.

32 Todas as citações sobre segurança de IA e a equipe de Superalignment foram retiradas de

uma gravação de áudio da reunião geral, em 15 de maio de 2024.

33 Jan Leike (@janleike), "Entrei porque achava que a OpenAI seria o melhor lugar do mundo para fazer esse tipo de pesquisa. No entanto, venho discordando da liderança da OpenAI sobre as prioridades centrais da empresa já há bastante tempo, até que finalmente chegamos a um ponto de ruptura", Twitter (atual X), 17 de maio de 2024, x.com/janleike/status/1791498178346549382.

34 Kelsey Piper, "ChatGPT Can Talk, but OpenAI Employees Sure Can't", *Vox*, 17 de maio de 2024, vox.com/future-perfect/2024/5/17/24158478/openai-departures-sam-altman-employees-chatgpt-release.

35 Daniel Kokotajlo escreveu extensamente sobre seu processo de tomada de decisão, inclusive neste fio: Daniel Kokotajlo (@DKokotajlo67142), 15/1: "Em abril, pedi demissão da OpenAI depois de perder a confiança de que a empresa agiria de forma responsável em sua tentativa de construir inteligência artificial geral — 'sistemas de IA que são, de modo geral, mais inteligentes do que os seres humanos", Twitter (atual X), 4 de junho de 2024, x.com/DKokotajlo67142/status/1797994238468407380; e em seus posts e comentários no fórum de segurança de IA LessWrong: "Daniel Kokotajlo", *LessWrong*, acessado em 25 de novembro de 2024, lesswrong.com/users/daniel-kokotajlo. O valor estimado de suas ações vem de Kevin Roose, "OpenAI Insiders Warn of a 'Reckless' Race for Dominance", *New York Times*, 4 de junho de 2023, nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html.

36 Todas as citações do Slack foram retiradas de capturas de tela.

37 Sam Altman (@sama), "sobre as coisas recentes a respeito de como a openai lida com ações: nunca ficamos com as ações adquiridas de alguém, nem faremos isso se as pessoas não assinarem uma cláusula de sigilo (ou não concordarem com um acordo de não difamação). ações adquiridas são ações adquiridas, ponto-final. havia uma cláusula sobre possível cancelamento de ações nos nossos antigos documentos de rescisão;

embora nunca tenhamos recuperado nada, isso nunca deveria ter constado em qualquer documento ou comunicação. a responsabilidade é minha e uma das poucas vezes em que fiquei genuinamente envergonhado à frente da openai; eu não sabia que isso estava acontecendo e deveria saber. a equipe já estava no processo de corrigir a papelada padrão de rescisão no último mês, mais ou menos. se algum ex-funcionário que assinou um desses acordos antigos estiver preocupado com isso, pode me contatar e resolveremos também. sinto muito por isso", Twitter (atual X), 18 de maio de 2024, x.com/sama/status/1791936857594581428.

38 Bobby Allyn (@BobbyAllyn), "Declaração de Scarlett Johansson sobre a situação com a OpenAI. Uau:", Twitter (atual X), 20 de maio de 2024, x.com/BobbyAllyn/status/1792679435701014908.

39 Kylie Robison, "ChatGPT Will Be Able to Talk to You Like Scarlett Johansson in *Her*", *The Verge*, 13 de maio de 2024, theverge.com/2024/5/13/24155652/chatgpt-voice-mode-gpt4o-upgrades.

40 Sarah Krouse et al., "Behind the Scenes of Scarlett Johansson's Battle with OpenAI".

41 OpenAI, "How the Voices for ChatGPT Were Chosen", *OpenAI* (blog), 19 de maio de 2024, openai.com/index/how-the-voices-for-chatgpt-were-chosen/.

42 Krouse et al., "Scarlett Johansson's Battle with OpenAI".

43 Allyn, "Statement from Scarlett Johansson".

44 OpenAI, "How the Voices for ChatGPT Were Chosen".

45 Derek Robertson, "Sam Altman's Scarlett Johansson Blunder Just Made AI a Harder Sell in DC", *Politico*, 22 de maio de 2024, politico.com/news/magazine/2024/05/22/scarlett-johansson-sam-altmans-washington-00159507.

46 Todas as descrições e citações da reunião são baseadas em uma gravação de áudio da reunião, em 22 de maio de 2024.

47 Kelsey Piper, "Leaked OpenAI Documents Reveal Aggressive Tactics Toward Former Employees", *Vox*, 22 de maio de 2024, vox.com/future-perfect/351132/openai-vested-equity-nda-sam-altman-documents-

employees.

48 Todas as citações são de uma gravação de áudio da reunião, em 23 de maio de 2024.

49 Deepa Seetharaman, "Turning OpenAI into a Real Business Is Tearing It Apart", *The Wall Street Journal*, 27 de setembro de 2024, wsj.com/tech/ai/open-ai-division-for-profit-da26c24b.

## CAPÍTULO 18: A FÓRMULA DO IMPÉRIO

1 Tyler Cowen, apresentador, *Conversations with Tyler*, podcast, episódio 61, "Sam Altman on Loving Community, Hating Coworking, and the Hunt for Talent", Mercatus Center Podcasts, 27 de fevereiro de 2019.

2 Sam Altman, "Successful People", *Sam Altman* (blog), 7 de março de 2013, blog.samaltman.com/successful-people.

3 Sam Altman, "Who Will Control the Future of AI?", Opinion, *Washington Post*, 25 de julho de 2024, washingtonpost.com/opinions/2024/07/25/sam-altman-ai-democracy-authoritarianism-future/.

4 Todas as citações da reunião geral foram retiradas de uma gravação de áudio, em 15 de maio de 2024.

5 Captura de tela de mensagem no Slack.

6 Amir Efrati e Wayne Ma, "OpenAI CEO Cements Control as He Secures Apple Deal", *The Information*, 29 de maio de 2024, theinformation.com/articles/openai-ceo-cements-control-as-he-secures-apple-deal.

7 Aaron Holmes, Natasha Mascarenhas e Julia Hornstein, "OpenAI CEO Says Company Could Become Benefit Corporation Akin to Rivals Anthropic, xAI", *The Information*, 14 de junho de 2024, theinformation.com/articles/openai-ceo-says-company-could-become-benefit-corporation-akin-to-rivals-anthropic-xai.

8 Kevin Roose, "OpenAI Insiders Warn of a 'Reckless' Race for Dominance", *New York Times*, 4 de junho de 2024, nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html.

9 "A Right to Warn About Advanced Artificial Intelligence", acessado em 5 de novembro de 2024, righttowarn.ai.

10 Pranshu Verma, Cat Zakrzewski e Nitasha

Tiku, "OpenAI Illegally Barred Staff from Airing Safety Risks, Say", *Washington Post*, 13 de julho de 2024, washingtonpost.com/technology/2024/07/13/openai-safety-risks-whistleblower-sec/.

11 Pranshu Verma, Cat Zakrzewski e Nitasha Tiku, "Senators Demand OpenAI Detail Efforts to Make Its AI Safe", *Washington Post*, 23 de julho de 2024, washingtonpost.com/technology/2024/07/23/openai-senate-democrats-ai-safe/.

12 OpenAI, "OpenAI Welcomes Sarah Friar (CFO) and Kevin Weil (CPO)", *OpenAI* (blog), 10 de junho de 2024, openai.com/index/openai-welcomes-cfo-cpo/.

13 John Schulman (@johnschulman2), "Compartilhei a seguinte nota com meus colegas da OpenAI hoje: tomei a difícil decisão de deixar a OpenAI. Essa escolha decorre do meu desejo de aprofundar meu foco no alinhamento da IA e iniciar um novo capítulo da minha carreira, no qual posso retornar ao trabalho técnico direto. Decidi perseguir esse objetivo na Anthropic, onde acredito que posso ganhar novas perspectivas e fazer pesquisas ao lado de pessoas profundamente envolvidas nos temas que mais me interessam…", Twitter (atual X), 5 de agosto de 2024, x.com/johnschulman2/status/1820610863499509855.

14 Greg Brockman (@gdb), "Vou tirar um sabático até o fim do ano. Primeira vez que descanso desde que fundamos a OpenAI há 9 anos. A missão está longe de terminar; ainda temos uma IAG segura para construir", Twitter (atual X), 5 de agosto de 2024, x.com/gdb/status/1820644694264791459.

15 Mira Murati (@miramurati), "Compartilhei a seguinte nota com a equipe da OpenAI hoje", Twitter (atual X), 25 de setembro de 2024, x.com/miramurati/status/1839025700009030027.

16 Bob McGrew (@bobmcgrewai), "Acabei de compartilhar isto com a OpenAI:", Twitter (atual X), 25 de setembro de 2024, x.com/bobmcgrewai/status/1839099787423134051; Barret Zoph (@barret_zoph), "Publiquei esta nota para a OpenAI", Twitter (atual X), 25 de setembro de 2024, x.com/barret_zoph/status/1839095143397515452.

17 OpenAI, "Introducing OpenAI o1", *OpenAI* (blog), acessado em 6 de janeiro de 2025, openai.com/o1.

18 Dara Kerr, "How Memphis Became a Battleground over Elon Musk's xAI Supercomputer", NPR, 11 de setembro de 2024, npr.org/2024/09/11/nx-s1-5088134/elon-musk-ai-xai-supercomputer-memphis-pollution.

19 Stephanie Palazzolo, Erin Woo e Amir Efrati, "How Anthropic Got Inside OpenAI's Head", *The Information*, 12 de dezembro de 2024, theinformation.com/articles/how-anthropic-got-inside-openais-head, "How Claude Became Tech Insiders' Chatbot of Choice", *New York Times*, 13 de dezembro de 2024, nytimes.com/2024/12/13/technology/claude-ai-anthropic.html.

20 Kenrick Cai, Krystal Hu e Anna Tong, "Exclusive: OpenAI Co-Founder Sutskever's New Safety-Focused AI Startup SSI Raises $1 Billion", *Reuters*, 4 de setembro de 2024, reuters.com/technology/artificial-intelligence/openai-co-founder-sutskevers-new-safety-focused-ai-startup-ssi-raises-1-billion-2024-09-04/.

21 Stephanie Palazzolo, Erin Woo e Amir Efrati, "OpenAI Shifts Strategy as Rate of 'GPT' AI Improvements Slows", *The Information*, 9 de novembro de 2024, theinformation.com/articles/openai-shifts-strategy-as-rate-of-gpt-ai-improvements-slows; Deepa Seetharaman, "The Next Great Leap in AI Is Behind Schedule and Crazy Expensive", *The Wall Street Journal*, 20 de dezzembro de 2024, wsj.com/tech/ai/openai-gpt5-orion-delays-639e7693.

22 OpenAI, "New Funding to Scale the Benefits of AI", *OpenAI* (blog), 2 de outubro de 2024, openai.com/index/scale-the-benefits-of-ai/.

23 Sam Altman, "The Intelligence Age", *Sam Altman* (blog), 23 de setembro de 2024, ia.samaltman.com.

24 Gravação de áudio da reunião, em 26 de setembro de 2024.

25 Sam Altman (@sama), "Acabo de publicar esta nota para a OpenAI: Oi, pessoal — Mira foi fundamental para o progresso e o crescimento da OpenAI nos últimos seis anos

e meio; ela foi um fator extremamente significativo em nossa evolução de um laboratório de pesquisa desconhecido para uma empresa importante. Quando Mira me informou esta manhã que estava saindo, fiquei triste, mas naturalmente apoio sua decisão. No último ano, ela vinha formando um sólido conjunto de líderes que continuará nosso progresso. Também quero compartilhar que Bob e Barret decidiram deixar a OpenAI. Mira, Bob e Barret tomaram essas decisões de forma independente e amigável, mas o timing da decisão de Mira fez com que fizesse sentido fazer tudo de uma vez agora, para que possamos trabalhar juntos para uma transição suave para a próxima geração de liderança", Twitter (atual X), 25 de setembro de 2024, x.com/sama/status/1839096160168063488.

26 Musk v. Altman, No. 4:24-cv-04722, CourtListener (N.D. Cal., 14 de novembro de 2024), ECF No. 32.

27 Jessica Toonkel, Keach Hagey e Meghan Bobrowsky, "Meta Urges Califórnia Attorney General to Stop OpenAI from Becoming For-Profit", *The Wall Street Journal*, 13 de dezembro de 2024, wsj.com/tech/ai/elon-musk-open-ai-lawsuit-response-c1f415f8.

28 OpenAI, "Why OpenAI's Structure Must Evolve to Advance Our Mission", *OpenAI* (blog), 27 de dezembro de 2024, openai.com/index/why-our-structure-must-evolve-to-advance-our-mission/.

29 Sam Altman, "Reflections", *Sam Altman* (blog), 5 de janeiro de 2025, blog.samaltman.com/reflections.

## EPÍLOGO: CASTELO DE CARTAS

1 Karen Hao, "A New Vision of Artificial Intelligence for the People", *MIT Technology Review*, 22 de abril de 2022, technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people.

2 Entrevistas da autora com Kathleen Siminyu em novembro de 2021; Michael e Caroline Running Wolf, em novembro de 2021; Kevin Scannell, em dezembro de 2021; Vukosi Marivate, em abril de 2023; Pelonomi Moiloa e Jade Abbott, em abril de 2023; Matteo Wong, "The AI Revolution Is Crushing Thousands of Languages", *The*

*Atlantic*, 12 de abril de 2024, theatlantic.com/technology/archive/2024/04/generative-ai-low-resource-languages/678042/.

3 "Kevin Scannell on 'Language from Below: Grassroots Efforts to Develop Language Technology for Minoritized Languages' 24.S96 Special Seminar: Linguistics & Social Justice", publicado em 17 de novembro de 2021 pelo MIT-Haiti Initiative, Facebook, 2 h 56 min 46 s, facebook.com/mithaiti/videos/1060463734714819; OpenAI, "GPT-4", OpenAI, 14 de março de 2023, openai.com/index/gpt-4-research.

4 Entrevistas da autora com Keoni Mahelona entre outubro e novembro de 2021; e Peter-Lucas Jones entre novembro e dezembro de 2021 e em janeiro de 2022.

5 Entrevistas da autora com Mahelona; Jones; Caleb Moses (cientista de dados que trabalhou no projeto, em novembro de 2021); e várias outras pessoas envolvidas no trabalho de preservação linguística da Te Hiku, entre novembro de 2021 e janeiro de 2022.

6 Entrevista com Mahelona em outubro de 2021.

7 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey e Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision", preprint, *arXiv*, 6 de dezembro de 2022, p. 1-2, arxiv.org/pdf/2212.04356.

8 Mozilla, "About DeepSpeech", *Mozilla GitHub*, acessado em 16 de dezembro de 2024, mozilla.github.io/deepspeech-playbook/DEEPSPEECH.html.

9 Entrevistas da autora com Timnit Gebru em agosto de 2024; e com Milagros Miceli em agosto de 2024.

10 "About Us", *Distributed AI Research*, acessado em 16 de dezembro de 2024, dair-institute.org/about.

11 "Research Philosophy", *Distributed AI Research*, acessado em 16 de dezembro de 2024, dair-institute.org/research-philosophy.

12 "Data Workers' Inquiry", *Data Workers' Inquiry*, acessado em 16 de dezembro de 2024, data-workers.org.

13 Oskarina Veronica Fuentes Anaya, "Life of a Latin American Data Worker", *Data Workers'*

*Inquiry*, acessado em 16 de dezembro de 2024, data-workers.org/oskarina; correspondência da autora com Fuentes, em julho de 2024.

14 Entrevista da autora com Mophat Okinyi em agosto de 2024.

15 "Our Story", *Techworker Community Africa*, acessado em 16 de dezembro de 2024, techworkercommunityafrica.org/About.html.

16 Mophat Okinyi, "Impact of Remotasks' Closure on Kenyan Workers", *Data Workers' Inquiry*, acessado em 16 de dezembro de 2024, data-workers.org/mophat.

17 Billy Perrigo, "Mophat Okinyi", *Time*, 5 de setembro de 2024, time.com/7012787/mophat-okinyi.

18 Entrevista da autora com Daniel Pena em maio de 2024.

19 Ria Kalluri, "The Values of Machine Learning", palestra de conferência, 9 de dezembro de 2019, publicada em 9 de dezembro de 2019 pela NIPS 2019, SlidesLive, 28 min 51 s, slideslive.com/38923453/the-values-of-machine-learning.

20 Entrevista da autora com Deborah Raji em agosto de 2024.

21 Entrevista da autora com Sasha Luccioni em agosto de 2024.

22 Joseph Weizenbaum, "ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine", *Communications of the ACM* 9, n. 1, janeiro de 1966, p. 36-45, doi.org/10.1145/365153.365168.

# ÍNDICE

## A


Abbeel, Pieter, 56, 124, 235
Abbott, Andy, 40
aceleracionismo eficaz (e/acc), 233
aceleracionistas (*boomers*), 233, 301, 310, 3787, 386, 392, 393
Acemoglu, Daron, 94
Achiam, Joshua, 395
aconselhamento psicológico, 212, 213
acordos de não difamação, 384
África do Sul, 110, 120
Agência Internacional de Energia, EUA, 272
Agnew, William, 107, 112, 167
água mineral. *ver* recursos hídricos
Aguirre, Blaise, 331
AI Insight Forums, 307
AI Now Institute, 304
AI Scientist, 187, 313, 319, 341, 366
*AI takeoff* ("decolagem da IA"), 232
Airbnb, 45, 49, 142, 156, 205, 358
Alameda Research, 230
Algorithmic Justice League, 167
algoritmos, 59, 63, 67, 365
*Algoritmos da opressão* (Noble), 168
Alibaba, 26, 165
Allen & Company, 74
Alphabet, 111, 449
AlphaFold, 305
AlphaGo, 66, 99
Altman, Annie, 51-53, 320-333, 345-348, 455
  alegações de abuso sexual de, 15, 51-53, 321-322, 328-329, 345-346,395
  apelos à família por ajuda financeira, 320,325-326
  artigo da revista *New York*, 320-321, 322-323, 326-327, 328-329, 345-346, 395
  primeiros anos de vida e educação de, 29, 322-323
  problemas de saúde física de, 323, 326-327
  problemas de saúde mental de, 52-53, 323-324, 325, 3332-333
Altman, Jack, 29, 30, 35, 36, 321, 325, 328-329
Altman, Jerold "Jerry", 38-39, 328-329

Altman, Max, 38-39, 320-322, 25
Altman, Sam
"Era da Inteligência", 29, 394
arrecadação de fundos, 50, 394
benefícios da IAG, 19, 400
ChatGPT, 265-267, 278, 347
comparações com Jobs, 14, 46
comportamento de liderança, 339-352, 353-356, 373-376
conflitos e desavenças na OpenAI, 154, 155, 156, 240, 241, 309-311, 386
crise de ações e, 371-373, 374-376
crise de Johansson, 373, 381-383
demissão e reintegração de, 13-23, 355-363
a investigação, 360-361, 366-367, 368, 382
depoimento ao congresso de, 297-299, 307
educação de, 39-40
estrutura de negócios da OpenAI, 24-25, 68-70, 73-74
fórmula do sucesso de, 43-45, 149
fundação da OpenAI, 23-24, 36-38, 54-56
GPT-3, 138-140, 276
GPT-4, 245-253, 340, 371-374, 380-382
ideologia do altruísmo eficaz e, 228-229
irmã Annie e, 51-53, 320-333
alegações de abuso sexual, 15, 51-53, 321-322, 328-329, 345-346,395
Loopt e, 41-47, 51, 75
missão da OpenAI, 390-392
MIT Technology Review e, 81, 92
na Y Combinator (YC), 33, 37, 41, 48, 316
nascimento e primeiros anos de, 38-40
paranoia de, 152-153
patrimônio líquido, 190
personalidade, 41-44, 326, 340
plano da empresa de chips de IA, 15, 368-369
plano de comercialização,156
políticas de, 50-51, 69
Projeto Manhattan, 152, 311-313
reformas do escritório Mayo, 81
relações com a mídia, 42-43, 373
retiro de outubro de 2022, 254
roteiro de pesquisa, 66, 190
Scallion, 369-371
sexualidade, 40
sobre Napoleão, 389-390
turnê mundial, 307, 330
visão para a OpenAI, 20, 89, 147-148, 260
altruísmo eficaz (AE), 63, 228-229, 314
alucinações, 118, 119, 218, 266, 351
Amazon, 49, 54, 147, 167

data centers, 272, 277, 283
Mechanical Turk, 197
ameaças internas, 154
amianto, 285
Amodei, Daniela, 62, 161, 163, 230
Amodei, Dario, 62-659
"O Divórcio", 162, 186, 233, 242
demissão de Altman, 357
*Dota 2*, 134, 150
fundação da OpenAI, 38, 56
GPT-2, 130-134
GPT-3, 138-140, 142, 150, 161
na Anthropic, 65, 121, 134, 162, 230
Nest, 139, 150, 156, 161, 244
riscos e segurança da IA, 62-65, 128-131, 132, 137, 138, 150-159, 161-163
Android, 105, 238
anotação de dados, 26, 196-198, 402-403
carros autônomos, 197-199, 210, 216, 219
Scale IA, 205-209
trabalhadores quenianos, 26, 29, 197-199, 205, 208-210, 213-215, 414
trabalhadores venezuelanos, 198-208
Antel, 288-289, 291
Anthropic, 17, 65, 121, 162-163, 230
agregação de Leike, 378
avaliação, 28
Claude, 260, 350, 370, 394-395
falência da FTX e, 256
Frontier Model Forum, 301, 304-305
fundação de, e "O Divórcio", 65, 162, 233
AP Bio, 245
APEC CEO Summit, 14
APIs (interfaces de programação de aplicativos), 156, *ver também APIs específicas*
*Apollo 11* (documentário), 312
Apollo, programa, 312
Appen, 142, 193, 198, 200-204
Apple, 39, 205, 327, 391
aprendizagem de máquina, 60, 62, 84, 100, 103
aprendizagem por reforço a partir de feedback humano (RLHF), 151, 160, 180, 215-224, 257, 266, 377
Arancibia, Alexandra, 282, 283, 293, 449
Arizona, 26, 277, 279, 285, 289
armas autônomas, 59
armas biológicas, 301, 304
Arrakis, 267, 365
arsênico, 280
arXiv, 26
Asimov, Isaac, 89
Atacama, deserto do, 269, 280-283, 445

Audiência do Comitê Judiciário do Senado, 297-299, 302-304
Authors Guild, 140
*Automata Studies* [Estudos sobre autômatos], 95, 424
autoritarismo, 78, 153, 199, 390
Azure AI, 74, 264

## B

*babbage*, 155
Babbage, Charles, 155
Baidu, 26, 28, 62, 165, 420
Banco do Vale do Silício, crise do em 2023, 50
Banco Mundial, 210
banidos, 223
Bankman-Fried, Samuel, 230
Beckham, David, 13
Bell Labs, 62
Bender, Emily M., 169-170, 172, 174, 253
  "Sobre os perigos dos papagaios aleatórios", 170-179
Bengio, Samy, 167, 171
Bengio, Yoshua, 111, 167
Bezos, Jeff, 49
Biden, Joe, 119, 318
Bing, 121, 305
Birhane, Abeba, 107, 112, 142, 143
Black Lives Matter, 158, 168, 173
Bolt, Usain, 43
bomba atômica, 313
bondade, 405
Books2, 140
Books3, 431
Boric Font, Gabriel, 292
Bostrom, Nick, 36-37, 63-648, 128
bot, taxa de, 203
bots de acompanhantes, 183
bots sexuais, 183
Boyd, Eric, 264
Brady, Tom, 231
Bridgewater Associates, 230
Brin, Sergey, 248
Brockman, Anna, 21, 255-256, 327
Brockman, Greg
  Altman e, 242, 361, 366
    demissão e reintegração, 13, 20-21, 361, 366
  conselho administrativo e, 240
  conselho administrativo e de supervisão, 317
  cultura e missão da OpenAI, 61-62, 90-92
  desligamento de, 393

*Dota 2*, 134, 150
entrevista com a autora, 81-87, 90-92, 276
estrutura de governança da OpenAI, 68-70
fundação da OpenAI, 28, 46, 47, 48, 49, 50, 52
GPT-4, 245-253, 340
infraestrutura de computação, 272, 274
Latitude, 185
liderança da OpenAI, 65-77, 89-92, 242-243
Omnicrisis, 386-387
plano de comercialização, 156
retiro de outubro de 2022, 254
roteiro de pesquisa, 66
Scallion, 369-371
Stripe, 49, 54, 62, 65, 80, 88-89
Brundage, Miles, 247, 249, 309, 379, 395
Buolamwini, Joy, 167
*Burning Man*, 44, 262
Burrell, Jenna, 99
Buschatzke, Tom, 279

## C

"caixa-preta", 113
câncer, 195, 280, 285, 290, 297, 368
capital de mercado, 18, 87, 91
capitalismo de vigilância, 107, 116, 139
capitalismo do desastre, 193-224
Carnegie Mellon University, 103, 177
Carr, Andrew, 375
carros autônomos, 197-199, 210, 216, 2196
anotação de dados, 26, 196-198, 402-403
Carter, Ashton, 51
cenário sobrevivencialista, 36
cenários extremos, 118
Center for AI Safety, 317
Center on Long-Term, 379
Centre for the Governance of AI, 316-317
Centro de Segurança e Tecnologia Emergente (CSET), 18, 303, 316, 350
cérebro humano, 67, 96, 100, 103, 115
Cerrillos, Chile, 285-287, 292-293
cetamina, 44, 51
CFPB. *ver* Consumer Financial Protection Bureau
chatbots, 28, 101, 117, 118, 219, 222
ELIZA, 101-102, 177, 408
GPT-3, 215-216
GPT-4, 242-247
LaMDA, 159, 253
Meena, 158-159

Tay, 159
ChatGPT, 257-260, 262-267
GPT-3.5 como base, 257
incentivo de inscrição, 265
lançamento, 14, 29, 65, 106, 126, 219, 221, 257, 259, 298
modo de voz, 369, 381
problema de alucinações,118-119, 266
tradição conexionista, 101
Chauvin, Derek, 158
Chen, Mark, 372, 395
Chesky, Brian, 50, 358
Chicago Boys (Chicago school of economics), 270, 292
Chile, 26, 269-279
data centers, 282-288, 292-295
extrativismo, 271, 290-296, 405
golpe de Estado em 1973, 270
protestos de 2019 a 2022, 287, 292
Chile, chegada dos espanhóis no, 269
China
chips de IA, 121, 300
desenvolvimento de IA, 109, 137, 152, 165, 195, 301, 303–4, 305, 307, 309–10, 297-307
vigilância generalizada, 108
Chuquicamata, desabamento da mina em 1957, 279
CIA (Central Intelligence Agency), 270, 316
cibersegurança, 119, 154, 184
ciência sem fronteiras, 304
Clarifai, 114, 238
Clark, Jack, 82, 88, 131, 132, 160-161, 307
Clarke, Arthur C., 62
Claude, 260, 350, 394-395
cláusula de reversão, 379, 383, 386
Climate Change AI, 84-85
CLIP, 235
clipes de papel, 36, 64
CloudFactory, 208-209, 214
cobre, 269, 271, 275, 279-281, 287
Codex, 188, 242, 247, 267, 314
código aberto (open-source), 118, 127, 136, 142, 187, 236, 274, 300, 390
código-fonte, 64-65
Cogito, 242
cognição, 115, 125
Cohere, 302
Coinbase, 142
Collard, Rosemary, 110
Colômbia, 26, 109, 200-201, 403
“colonialismo de dados”, 109
Colorado, rio, uso da água, 278

combate ao abuso sexual, 237, 265
Comissão de Valores Mobiliários (SEC), EUA, 374-375, 393
Comissão Europeia, 111
Comissão Federal de Comércio (FTC), EUA, 238, 304, 351
Comitê de Segurança de Implementação (CSI), 318, 340, 344, 354, 355
Common Crawl, 140-142, 169
compressão, 127
computação quântica, 147
computadores. *ver compute*
*compute*, 66-68
Conception, 50
conexionismo, 100, 102-104, 115, 124
Conneau, Alexis, 369
consciência humana, 117, 125
Consumer Financial Protection Bureau (CFPB), 407
conteúdo pornográfico, 114, 193, 237-238, *ver também* material de abuso sexual infantil
controle humano, evasão de IA, 157, 306, 309, 370
Copilot, 188, 236, 238, 263, 329
Coreia do Norte, 152
Coreia do Sul, 66
corrida armamentista, 27
Couldry, Nick, 109
Cowen, Tyler, 389
Crab Generation, 222
Creative Commons, 187
crescimento econômico, 47-48
*Criadores de gênios* (Metz), 87
criogenia, 191
criptomoedas, 70, 190
"crowdwork anônimo", modelo de, 209
CSET. *ver* Centro de Segurança e Tecnologia Emergente
cubo mágico, 77
CUDA (Compute Unified Device Architecture), 68
*curie*, 155
Curie, Marie, 155
Curry, Stephen, 231
Cyc, 102

## D

D'Angelo, Adam, 316
  comportamento de liderança de Altman, 339-352, 353-356, 373-376
  demissão de Altman, 13-23, 355-363
Da Vinci, Leonardo, 155
dados vazados, 298
DAIR (Instituto de Pesquisa em IA Distribuída), 401-402, 405
Dalí, Salvador, 234
DALL-E, 23, 116, 120, 234-241, 257, 267, 310

Damon, Matt, 313
data centers, 26, 271-275
  emissões de carbono, 81-82, 165-166, 176-178
  Google, 272-273, 282-288, 291-292
  no Chile, 282-288, 292-295
  no Uruguai, 288-295
  uso de energia, 84, 86-87, 272-275, 277-279, 288-289
Data Workers' Inquiry (Investigação dos Trabalhadores de Dados), 402-403
Dataluna, 286
*davinci*, 155
Dean, Jeff, 35, 164, 166-167, 175
deep learning, 103-107, 110-117
  ImageNet, 55, 67, 106, 123, 258
  impactos discriminatórios de, 64, 114
  limitações e riscos de, 112, 114-115, 117
deepfakes, 86, 239, 382
DeepMind, 11, 17, 28, 33-35, 55, 66, 73, 258, 375
  AlphaFold, 305
  AlphaGo, 66, 99
  escala, 138, 164
  OpenAI e ChatGPT, 119, 125, 138, 160, 260
demonstração a Gates, 78, 138, 245
Dempsey, Jessica, 110
Departamento de Comércio dos EUA, 299-300, 303-304
Desafio do Manequim, 108
descaroçadora de algodão, 94
desigualdade, 194, 271, 287
desinformação, 59, 78, 183, 241, 368
Desmond-Hellmann, Sue, 366
destilação, 182, 302
Desvio, o, 365, 368, 374, 386, 387
Díaz Bejarano, Nicolás, 293
difusão, 235-236, 365
  Stable Diffusion, 120, 143, 236, 241, 281
Digital Realty, 272
direitos humanos, 30, 201
discurso de ódio, 29, 196, 210
dissonância cognitiva, 227
divisão norte-sul, 27, *ver também países específicos*
Divórcio, o, 162, 186, 230, 233, 242
DNNresearch, 55, 58, 104-105
Dorador, Cristina, 281
*Dota 2*, 73-74, 77, 134, 150, 244
*Doutor Estranho* (filme), 299
Dowling, Steve, 160, 255, 373
drones de vigilância, 59
Du, Yilun, 127

DUST, 267

## E

Edison, Thomas, 61-62
educação, 407-408
efeitos de network, 48-49, 191
Effective Ventures Foundation, 316
eleição de 2016, EUA, 47, 59
ELIZA, 101-102, 117, 408
emissões de carbono, 86, 165-166, 176-178, 273-274, 291
energia renovável, 84, 273-274
"engenheiro 10x", 88, 147, 148, 180, 182, 229
Enigma, 97
Equinix, 272
"Era da Inteligência", 29, 394
escalabilidade, 121, 138, 155, 162, 165, 181-183
  "leis de escala do ódio", 142
escravidão, 94, 211, 389
Escritório Mayo, 81, 311, 423
especismo, 34
*Estallido Social*, 287, 292
estrutura de lucro limitado, 77, 82, 316, 391
*estudos sobre autômatos*, 95
Etcheverry, Aisén, 296
Evolved Transformer, 166, 176-177
exploração do trabalho, 48, 139, 194, 197, 214, 410, *ver* anotação de dados
extinção, 34, 63, 239, 368
extrativismo, 109, 405
  no Chile, 269-288
  no Uruguai, 288-292
  uso do termo, 110

## F

Facebook, 22, 26-27, 47, 57, 59, 111, 160, 165, 195, 211, 222, 230, 238, 316, 327
Fact Factory, 259
Fairwork, 205, 208, 403
Fedus, Liam, 246, 395
Feynman, Richard, 127
filtragem de dados, 142, 182
Flamingo Generation, 222
Floyd, George, 158
Fórmula 1, 13, 231
Founders Fund, 47
Foursquare, 42
fraudes, 223, 265
Friar, Sarah, 393
Fridman, Lex, 373
Friedman, Milton, 270

Friend, Tad, 36, 40
*friendly AI* (IA amigável), 64, 314
Frontier Model Forum, 301, 304-305
FTX, 230-231
  falência, 232, 256, 317
FTX Future Fund, 231
Fuentes Anaya, Oskarina Veronica, 200, 402
Fundação Mozilla, 107, 400
fusão nuclear, 147, 190, 278
Futures of Artificial Intelligence Research [Futuros da Pesquisa em Inteligência Artificial], 271
"Futuro Perfeito", 378

## G

"ganhar para doar", 237
gargalo, 85, 245, 277
Gates, Bill, 75
  demonstração da OpenAI, 78, 138, 245
  depoimento no Congresso de, 307
  GPT-4, 245
Gawker Media, 47
"geração composicional", 245
GDPR (General Data Protection Regulation), 141
Gebru, Timnit, 34, 60-62, 114, 166-177, 273, 401-402
Geometric Intelligence, 116
geração de código, 157, 181, 186-188, 313
geração de texto, 129, *ver também* previsão da próxima palavra
*Ghost Work* (Gray e Suri), 196
Gibstine, Connie, 38-40, 53, 321
Gibstine, Marvin, 38
GitHub, 141, 186-189, 236, 242, 329
  Codex, 188, 242, 247, 267, 314
  Copilot, 188, 236, 238, 263, 329
GiveWell, 230, 317
Gmail, 106
Go (jogo), 66
Gobi, 267, 342
Goldman Sachs, 29, 273
Good Ventures, 230
Google, 11, 126
  "revolta contra a tecnologia", 59
  Amodei no, 62
  Android, 106, 238
  avaliação, 77
  captchas, 104
  data centers, 272-273, 282-288, 291-292
  DeepMind. *ver* DeepMind
  DNNresearch, 55, 58, 104-105

Frontier Model Forum, 301, 304-305
GPT-4 e, 249
LaMDA, 159, 253
modelo Imagen, 241
pesquisa de IA, 71, 76, 78, 106, 112, 183
Projeto Maven, 59
vasculhagem de IA, 140
reconhecimento de fala, 98
redes neurais, 106-107
Sutskever e, 54, 106
Transformers, 126-127, 129, 138, 234-235, 302
Waymo, 106
Google Brain, 784, 164, 171, 173
Google I/O, 370-371, 374
Google Research, 60, 164, 168, 171
Google Tradutor, 106, 126, 201, 398
Gordon-Levitt, Joseph, 318
GPT-1, 183
lançamento, 27, 128
treinamento e capacidades, 128-130, 235
GPT-2, 78, 252
demonstração a Gates, 78, 138
erros, 151-152
escala,136-137
hipótese da "linguagem pura", 135-136
lançamento, 78, 131, 310
retenção de pesquisas, 131-133, 135, 171
riscos potenciais, 131-132
treinamento e capacidades, 130, 134, 155, 158, 398
GPT-3, 138-140, 276, 289
API, 156-162, 164, 167-170, 219, 310
imitação de chatbot, 118
InstructGPT, 216-217, 219, 246
lançamento, 139, 158
treinamento e capacidades, 138-140, 156-157, 186, 244, 276
GPT-3.5, 257-258, 267
GPT-3.75, 369
GPT-4, 193, 242-253
Bing, 117-119, 247
capacidades, 27, 119, 141, 244-245
desenvolvimento, 245-251
Superassistant, 246-249, 256-257, 372
GPT-4 Turbo, 340, 355
GPT-4o, 371-372, 374, 376, 381, 391
GPT-5, 277, 319, 365
Orion, 365, 370-371, 394
GPUs (graphics processing units), 68, 138, 140, *ver também* Nvidia

escassez, 259
Graham, Paul, 38, 41, 45, 47, 49, 75
grandes modelos de linguagem, 26, 79, 121, 129, 136, 150, 159, 161, 164-165, 168-170
"Sobre os perigos dos papagaios aleatórios",170-179, 253, 273, 401
perda de línguas, 418
Gray, Mary L., 196-197
Groom, Lachy, 50
Groves, Leslie R., 313
Guo, Eileen, 190

## H

*Hacker News*, 77
hackers do bem, 113
Hanna, Alex, 401
*hardware, excedente de*, 181
Harris, Kamala, 298
Hassabis, Demis, 33-35, 55, 305
Helion Energy, 190, 278
Hendrycks, Dan, 317
Hepburn, Audrey, 101
*Her* (filme), 246, 369, 372-373, 381, 383
Herbert-Voss, Ari, 184-185
Hernández, Andrea Paola, 206, 414
Hernandez, Danny, 67
Herzberg, Elaine, 113, 118
Hinton, Geoffrey, 103, 111
DNNresearch, 55, 58, 104-105
ImageNet, 55, 67, 106, 103, 106, 123, 258
redes neurais e deep learning, 103-107, 110-117
Sutskever e, 55, 103, 105-106, 114, 116, 123, 127, 254, 258
hiperescaladores, 272, 282, 290, 292
hipótese da "linguagem pura", 135-136, 164, 234, 313
hipótese do "fundamento", 135, 313
Ho, Jonathan, 235
Hoffman, Reid, 58, 70, 315, 358, 375
Hogan, Mél, 272
Hollywood, 298
Hood, Amy, 79
Hooker, Sara, 302-303, 306
Huffman, Steve, 43
Huggines, Ricardo, 207
Hugging Face, 274, 407
Hurd, Will, 316
Huyen, Chip, 60
Hydrazine Capital, 44, 47, 49, 75

## I

IA (inteligência artificial)

antropomorfização, 96, 117
benefícios, 23-24, 27, 30, 83-86, 94-96, 327, 390, 405
cobertura da autora, 26-27
comercialização, 26-27, 58-59, 82, 106-120, 156
comparação com a IAG, 78
convenções de artigos e revisões por pares, 26
definição de inteligência, 95-99
êxodo de docentes de pesquisa, 111
financiamento, 106-112, 116, 156
impérios de, 27-28, 30, 200, 223-224, 332-333, 401, 405, 407
investimentos da iniciativa privada, 110
linha do tempo, 98, 138-139, 232, 259, 378-380
pesquisa e desenvolvimento, 25-26, 28-29, 106-112
políticas regulatórias, 35, 141, 264, 270, 295, 297, 299-307, 374
raspagem/extração/coleta de dados, 107, 110, 119, 141-143, 399
riscos e danos de, 27-30, 33-35, 62-65, 84-87, 112-116
teorias de, 99-106
treinamento de modelos, 16, 68, 103, 139-143, 169, 243-245, 276-279, 302
uso do termo, 95-97, 390
IA descontrolada, 63, 64, 151, 231, 314
IA em escala cerebral, 67
IA neurossimbólica, 115, 121
IA, padronização de, 36, 242
desalinhamento, 63, 151, 315
Superalignment, 311, 346, 377-378
IA, segurança de, 62-65, 128-138, 297-307, 311-319, 406, *ver também* privacidade de dados; riscos existenciais
alinhamento e, 128, 150-151, 311
altruísmo eficaz e, 63, 228-229, 314
Audiência do Comitê Judiciário do Senado, 298-304
Frontier Model Forum, 301, 304-305, 309-311
limites, 298-304
IAG (inteligência artificial geral), 55-56, 82-86, 134-138, 232, 378-380
Google e, 34-35
OpenAI e Altman, 18-19, 23-24, 29, 41, 55-57, 69, 72-74, 117, 127, 148-149, 187, 240, 253-254, 297, 315, 349, 390-391
uso do termo, 83-84, 99-100
IBM, 104-105, 167
Watson, 104
Imagen, 241
ImageNet, 55, 67, 106, 123, 258
impacto ambiental, 86, 91, 176, 176-177, 291, 407
terra saqueada, 269-296
impacto discriminatório, 59-60, 64, 114, 120,142, 167-170, 184,305-306
impactos no local de trabalho, 84-85, 120, 222, 263-264
império em dissolução, 409-410
impérios da IA, 27-28, 30, 200, 223-224, 332-333, 401, 405, 407
implementação iterativa, 310, 370, 391

incentivo de inscrição, 265
Index Ventures, 206
Índia, 138, 195, 205, 242, 273, 318
Índice de IA, 110
industrialização, 48, 270-271
inferência, 104, 265-265, 364, 369
Inflection AI, 315, 375
*Information, The*, 42, 215, 277, 362, 365, 392
Inglewood, 277
Instacart, 354, 366
Instagram, 327, 393
InstructGPT, 216-217, 219, 246
inteligência, 115
  definição de, 95-99
inteligência artificial. *ver* IA
inteligência artificial geral. *ver* IAG
inteligência humana. *ver* inteligência
"inverno da IA", 103, 425
Irving, Geoffrey, 164, 361
Isaac, William, 109
Israel, 54, 209, 330

## J

Janah, Leila, 195, 209
*Jeopardy!* (programa de televisão), 104, 222
Jernite, Yacine, 274, 305
Jobs, Laurene Powell, 14
Jobs, Steve, 14, 45-46
*jogo da imitação, O* (filme), 88, 97
Johansson, Scarlett, 373, 381-383
John Burroughs School, 39
Johnson, Josh, 372
Johnson, Simon, 94
Jones, Peter-Lucas, 398
Jones, Shane, 238
Jonze, Spike, 246, 372
Jordan, Michael, 43

## K

Kacholia, Megan, 171-175
*kaitiakitanga*, 399
Kalluri, Ria, 107, 112, 405-406, 413
Kaplan, Jared, 162
Karnofsky, Holden, 63, 65, 230, 316-317
Karpathy, Andrej, 70-71
Kay, Alan, 316
Kelton, Fraser, 156, 236, 241, 246
Kennedy, John F., 61

Kennedy, John Neely, 297
Khan Academy, 245
Khan, Sal, 245
Khlaaf, Heidy, 184
Khosla Ventures, 77
Klein, Ezra, 121
Klein, Naomi, 270
Knight, Will, 131
Koko (gorila), 253
Kokotajlo, Daniel, 379-380, 384, 393
Kolln, Ryan, 142, 193
Krisiloff, Matt, 50
Krizhevsky, Alex, 55, 105-106, 123, 258
Kwon, Jason, 19-20, 340, 357, 364, 382-386

## L

Lago Vitória, 209
LAION, 142
LaMDA, 159, 253
Latitude, 185, 193
Lattice, 44
Leap Motion, 76, 156, 338
LeCun, Yann, 111, 165, 235, 300
Lei da OpenAI, 67-68, 91, 122, 129
Lei de Moore, 67-68, 121
Leike, Jan, 377-378
  alinhamento e segurança, 247, 249, 309, 311, 393
  desligamento de, 377-378, 391
"leis de escala do ódio", 142
leis de escalabilidade, 121, 155, 161-162, 165, 181-183, 302
Lemoine, Blake, 253
Lessin, Jessica, 42-43
liberdade de expressão, 360
Library Genesis, 140
Lightcap, Brad, 16-18, 76, 364, 384
língua de sinais dos Estados Unidos, 253
LinkedIn, 58, 220
LISTSERV, 36, 168, 173
lítio, 269, 271, 275, 280-281, 287
Liu Cixin, 89
Livingston, Jessica, 41, 46, 58, 75
Llama, 301
longevidade humana, 190
Loopt, 41-47, 51, 75
*Loopt Star*, 42
Lourd, Bryan, 373, 381
Lovelace, Ada, 155

Luccioni, Sasha, 273-274, 305, 407
Luka, Inc., 184
luo, povo, 209-210
Lydic, Desi, 372
Lyft, 205, 324

## M

MacAskill, William, 229-231
Machine Learning for Health, 84
Mądry, Aleksander, 18-19, 358, 370, 383, 388, 393
Maduro, Nicolás, 199
Mahelona, Keoni, 398-400
Makanju, Anna, 19, 160, 255, 298, 357
Mallery, Rob, 262
maori, povo, 397-399
Máquina de Turing, 88, 95
Marcus, Gary, 115, 124, 251, 298
Marte, 33-34, 282
Martin, George R. R., 140
material de abuso sexual infantil, 142-143, 185, 193, 196, 211, 237, 241
Mathenge, Richard, 403
Mayer, Katie, 156
McCarthy, John, 95-97, 390
McCauley, Tasha, 316-319, 366
  comportamento de liderança de Altman, 319, 345, 349-352, 353-354
  demissão de Altman, 18, 22
McGrew, Bob, 76, 161, 236, 243, 364, 394-395
Mechanical Turk (MTurk), 197-198, 205
Meena, 158
*megacampus*, 272
mega-hiperescalabilidade, 273
Mejias, Ulises A., 109
meritocracia, 45
Messerschmidt, Neily, 328-329
Meta, 59
  data centers, 271-272, 278, 280
  investimentos em IA, 110
  Llama, 301
  moderação de conteúdo, 194-196, 211
  OpenAI e, 165, 395-396
  open-source, 299-300
  recursos computacionais, 301
  revolta contra a tecnologia, 59
  Threads, 259
Metz, Cade, 87, 96
Metz, Luke, 246, 395
Miceli, Milagros, 402

Michelangelo, 87
Microsoft
  Altman e, 342-344
   demissão, 20-21, 358
  Azure AI, 74, 264
  Bing, 117-119, 247, 262-263, 348
  capitalização de mercado, 28, 87, 91, 289
  Copilot, 188, 236, 238, 263, 329
  data centers, 255, 271-279, 282-283, 292-295
  Frontier Model Forum, 301, 304-305
  GitHub, 141, 186-189, 236, 242, 329
  Helion Energy, 190, 278
  Inflection AI, 315, 375
  Max Altman na, 45
  reconhecimento de voz, 105
  ResNet, 305
  Tay, 159
Microsoft e OpenAI, parceria entre, 18, 75-76, 80-81, 240, 269-276, 398
  ChatGPT, 257-260, 262-267
  fases de cálculo, 282-285
  GPT-3, 138-140, 276, 289
  GPT-4, 193, 242-253
  investimentos e financiamento, 25, 27, 81, 84, 89, 92-93, 141-142, 53-155, 167, 254, 332
Microsoft Office, 247, 263
Microsoft Research, 74
Microsoft Teams, 263
Mighty AI, 198
Millicent, 220-224
Minsky, Marvin, 100-103
Mishra, Nikhil, 65, 253
Mission District, 64, 80
MIT (Massachusetts Institute of Technology), 18, 54, 94, 101, 111
*MIT Technology Review*, 11, 81, 92, 131, 174, 190, 361
Mitchell, Margaret "Meg", 168, 170, 172, 174, 253
modelos de fronteira, 301-307
modelos esparsos, 182
modelos multimodais, 98, 181, 233-234, 366
moderação de conteúdo, 142-143, 160, 184-185, 193-196, 200, 209-214, 250, *ver também* anotação de dados
  Sama, 194-196, 209-215, 220, 242, 403-404
Moeroa, Raiha, 398
Mohamed, Shakir, 109
monopólios, 48, 107, 147-148, 187, 299
Montgomery, Christina, 303
Moore, Gordon, 61, 121
Morton, Samuel, 97
Mosacat, 285-290, 293, 295, 405
Moskovitz, Dustin, 230

mudança climática, 34, 60, 83-86, 170, 200, 273, 278, 289, 297
Mundie, Craig, 75
Murati, Mira
  Altman e, 243, 346-352
    CEO interina, 20, 349, 356-357
    comportamento de liderança, 339-352, 353-356, 373-376
    demissão e reintegração, 13-23, 353-362
  contratação de, 76, 337
  DALL-E e, 240
  desligamento de, 394-395
  diretora de tecnologia, 337, 339
  histórico de, 76, 337-339
  Johansson e as crises de ações, 383
  Microsoft e, 186, 189, 268
  na Tesla, 338, 354
  Omnicrisis, 386-387
  Scallion, 369-371
  Superalignment, 377-378
  Toner e, 337, 342-344, 348-350
Murphy, Cillian, 313
Musk, Elon
  Altman e, 33-36, 69-70, 153, 312, 373
    comportamento de liderança, 353-356
    demissão, 353-362
  depoimento ao Congresso de, 307
  desligamento da OpenAI, 66-68
  estrutura de governança da OpenAI, 68-70
  financiamento, 68-74
  fundação da OpenAI, 23-24, 36-38, 54-62
  MIT Technology Review, tuítes e, 92
  Neuralink, 70, 80, 153, 315
  Page e, 34-35, 38, 248
  Projeto Manhattan, 311-313
  Radford e, 127
  riscos da AI, 33-35
  SpaceX, 33, 35, 37, 58, 360
  xAI, 316-317, 387, 392, 394
  Zilis e, 315, 319
  Zuckerberg e, 396
Mutemi, Mercy, 214, 288

## N

Nadella, Satya, 119
  demissão de Altman, 16, 18, 21, 358
  depoimento no Congresso de, 319
  GPT-4, 254, 255, 307
  parceria na OpenAI, 74-79, 247, 263

Nairóbi, Quênia, 194-196, 209-210, 213, 220, 403, 412
Napoleão Bonaparte, 389-390
Nectome, 191
negros na IA, 60-61, 166
Nepal, 208
Nest, 139-141, 150-151, 156-157, 162, 244
Netflix, 66, 77
Neural Architecture Search, 166, 176, 178
Neuralink, 70, 80, 153, 315
NeurIPS (Neural Information Processing Systems), 405
  Climate Change AI, 84
  Gebru e racismo, 60, 166
  OpenAI na, 159, 258, 364
  Test of Time, prêmio, 258, 364
New Enterprise Associates, 41
*New York* (revista), 320, 329, 331, 337, 345
*New York Times, The*, 87, 96, 101, 117, 121, 222, 244, 262, 268, 270, 298, 308, 359, 362, 374, 390
*New Yorker, The*, 34, 36, 40, 64
Newsom, Gavin, 306
Nkosi, Thami, 110
Noah, Trevor, 25
Noble, Safiya Umoja, 168
Norte da África, 208
Nova Zelândia, 36, 397-399, 413
Nvidia, 68, 276, 300, 399
  A100s, 180-181, 236, 242, 276-277
  B100s, 277
  H100s, 277, 279
  V100s, 138, 276

## O

Obama, Barack, 35, 51, 160, 210
Odisseu, 277
Okinyi, Albert, 211, 213-214
Okinyi, Cynthia, 210-213
Okinyi, Mophat, 196, 209, 213-214, 402, 414
Olin College of Engineering, 127, 399
Olson, Parmy, 29
Ommer, Björn, 236
Omni, 371-372
Omnicrisis, 380, 386-387, 390-392, 394
Open Philanthropy, 63, 230, 317
OpenAI. *ver também pessoas e produtos específicos*
  "mudança de paradigma", 142, 193, 214
  ações e crise de ações, 75-77, 378-380, 382-386
  avaliação, 14, 23, 25, 29, 57, 77, 90, 315, 395
  comercialização, 23-25, 73-74, 78-79, 107, 116, 156-160, 180, 265, 392

conflitos e rupturas na empresa, 150-155, 161-163, 233-234, 239-242, 265-267, 339-345, 377, 386, 392-394
conselho administrativo. *ver* OpenAI, conselho administrativo da
crise de Johansson, 371, 381-383
cultura organizacional, 61-62, 132, 152, 162-163, 260-265
demissão e reintegração de Altman, 13-23, 355-363
investigação, 360-361, 366-367, 368, 382
reportagem da autora, 23, 361-363
desligamento de Musk, 66-68
divisão de Pesquisa, 74, 79, 186, 239-240, 259, 266-267, 276, 309, 311, 313, 341, 358, 365-366
AI Scientist, 187, 313, 319, 341, 366
entrevistas da autora, 82-93
esforços de recrutamento, 56, 58-59, 65, 159
estrutura e governança empresarial, 24-25, 68-70, 73-74
com fins lucrativos, 25, 69-73, 265, 361-363, 392
lucro limitado, 77, 82, 316, 391
reestruturações de Altman, 93, 392
sem fins lucrativos, 17, 24-25, 37-38, 57-58, 69-77, 237-239, 252, 262
sociedade limitada, 73, 75-77
expansão, 73, 123-131, 152, 165-166, 180-183, 215-216, 242, 276-277, 303, 364-365, 395
financiamento, 68-74, 77-79, 137, 147, 161, 260-261, 315-316, 324, 358, 368, 394
Frontier Model Forum, 301, 304-305
funcionários, 255, 260-262, 375
fundação da, 23-24, 36-38, 54-62
jantar no Rosewood Hotel, 37, 54-56, 62
gargalo de dados, 245, 277
IA generativa e, 116-120
lançamento da, 58-61
logotipo, 15, 89, 375
missão da empresa, 73, 77, 239, 372
nomeação da, 24
Desvio, o, 365, 368, 374, 386, 387
Divórcio, o, 162, 186, 230, 233, 242
oferta secundária, 14, 16, 28, 359
parceria com a Microsoft. *ver* Microsoft e OpenAI, parceria entre
Pesquisa Exploratória, 155-157
projeto de edifícios e escritórios, 80-81, 311
remuneração, 58, 70-71, 76-77
reportagem da autora, 23-27
retiro de outubro de 2022, 254
roteiro de pesquisa, 66, 190, 232, 242, 276
segurança, 150-166, 247-267, 372-376, 368-373, 377-380, 382-385, 392-393
subdivisão da área de produtos, 156-161, 183, 215, 236
visão de Altman para, 20, 89, 148-149, 260-261
OpenAI Startup Fund, 191-192, 354
OpenAI, conselho administrativo da
comportamento de liderança de Altman, 339-352, 353-356, 373-376
carta aberta, 21-22, 359

Murati como CEO interina, 20, 349, 356-357
reportagem da autora, 361-363
membros se desligando e se integrando, 22, 64-65, 315-317, 365-366
questões de supervisão, 317-319
OpenResearch, 189-190
*Oppenheimer* (filme), 311-313
Oppenheimer, J. Robert, 312-313
ordem executiva 14110, 305
Orion, 365, 370-371, 394
Ortiz, Karla, 299
Ostrich, 222
Otero Verzier, Marina, 293-295
Oxford Internet Institute, 205, 403
*p(doom)*, probabilidade de destruição, 232, 249

## P

Pachocki, Jakub, 150
AI Scientist, 313-314, 341
Altman e, 314, 376-377
*Dota 2*, 150, 244
GPT-3, 244-245
GPT-4, 244-245, 308
novo cientista-chefe, 376
Omnicrisis, 395
segurança de IA, 150, 154
Page, Larry, 34-35, 38, 248
pandemia de covid-19, 62, 81, 155, 158, 186, 195, 206-207, 209-210, 215, 220, 231, 289, 311, 318, 325, 413
Pang, Wilson, 202
"pântanos de dados", 142
Paquistão, 223
paradigma, mudanças de, 142, 193, 214
Parakhin, Mikhail, 348
Park, Matt, 207
Parque de las Ciencias, 288
Partido Democrata, EUA, 50, 231
Patterson, Dave, 177
PayPal, 46, 58, 147, 201
PBJ1, PBJ2, PBJ3 e PBJ4, 196
pecuária, 229
Pena, Daniel, 290-291, 293, 404
Perceptron, 96, 100-101
*Perceptrons* (Minsky), 100, 102
perda de línguas, 398
perguntas de domínio fechado, 266
Perrigo, Billy, 143, 195, 212
pesos dos modelos, 153-154, 157, 264, 301
Pesquisa Exploratória, 155-157

pessimistas (*doomers*), 233, 301-302, 310, 363, 368, 378-379, 386, 392-393
Phillips Exeter, internato, 316
Phoenix, 277-278
Pichai, Sundar, 174, 307
Picoult, Jodi, 140
Pinochet, Augusto, 270, 292
Pioneer Building, 81, 311, 387
Piper, Kelsey, 378-380, 384, 393
Png, Marie-Therese, 109
pobreza, 195, 204, 209-210, 220, 279-280, 289, 327, *ver também* desigualdade
*poderoso chefão, O* (filme), 360
Poe, 319
políticas públicas, 30, 40, 131, 298, 307-308, *ver também* regulamentação
poltrona de abacate, 235, 237
poluição da água, 290
poluição do ar, 283
poluição sonora, 272
pontos cegos, 94
pornografia. *ver* conteúdo pornográfico
Posada, Julian, 199-200, 288
Prêmio Turing, 111, 167
preocupações com a privacidade. *ver* privacidade de dados
previsão da próxima palavra, 127, 129, 136
privacidade de dados, 30, 42, 64, 109, 190, 297, 303, 403
*problema dos três corpos, O* (Liu Cixin), 89
produtividade, 27, 29, 88, 120, 157, 160, 223, 243, 264
Projeto Chile, 270
Projeto de Pesquisa de Verão de Dartmouth (1956), 95, 100
Projeto Manhattan, 37, 152, 312-313
Projeto Manhattan de Alinhamento, 311
Projeto Maven, 59
proteção de denunciantes, 393

## Q

Q*, 364-365
QBRN, armas, 306, 370
Quadro de Preparação para Riscos (Preparedness Framework), 370
Queer in AI, 167, 405
Quênia, 143, 184, 194-196, 205, 327
    anotação de dados, 26, 29, 194-196, 208-210, 214, 219, 220, 402-403
    projetos de RLHF, 220, 223
Quilicura, Chile, 282-285, 287, 292-295
Quora, 18, 188, 316, 319

## R

racismo, 27, 97, 120, 159, 168, 170, 281
Radford, Alec, 127-129, 132, 140-142, 244
    CLIP, 235

desligamento de, 395
GPT-1, 127-129, 244
GPT-2, 132
Raji, Deborah, 63-64, 114, 167, 238, 302, 306-307, 406-407
Ralston, Geoff, 43, 45, 148
Ramesh, Aditya, 235
Ramos, Sonia, 279, 281-282, 291
rápida generalização, 159
raspagem de dados, 107, 110, 119, 141-143, 399
rastreamento de localização, 41-42
Raven, 277
reconhecimento de voz, 84, 105, 108, 123, 244, 305, 399-400
Whisper, 244, 246, 265, 400
reconhecimento facial, 64, 108, 110, 120, 167
recursos hídricos, 26, 279
*red teaming*, 184, 371
Reddit, 43, 140, 157, 169
redes neurais, 100-101, 103-105
alucinações, 118-119, 218, 266, 351
limitações e riscos, 112-116, 118-122
redes neurais densas, 182
redistribuição do poder, 406
refugiados sírios, 143
Regalado, Antonio, 190-191
regras governamentais. *ver* regulamentação
regulamentação (políticas regulatórias), 37, 91, 139, 141, 270, 295, 297, 299, 301, 307, 373, 410
Remotasks, 206-207, 220-224, 403
Renaldi, Adi, 190
renda básica universal, 92, 189
ResNet, 305
Retro Biosciences, 190
retropropagação, 103
revisão por pares, 26, 175, 365
revolução industrial. *ver* revoluções tecnológicas
revoluções tecnológicas, 13, 95, 99, 103
impérios da IA, 27-28, 30, 200, 223-224, 332-333, 401, 405, 407
*Rick e Morty* (desenho animado), 75
Rickover, Hyman G., 149
Rihanna, 13
risco de aceleração, 232, 248
riscos de informação, 131
riscos desconhecidos, 249
riscos existenciais, 34, 64, 131, 136, 151, 229-231, 249-250, *ver também* pessimistas (*doomers*)
*p(doom)*, probabilidade de destruição, 232, 249
Roberts Companies, 39
Robinson, David, 350
Roble, James, 322

robótica, 66, 73, 76-77, 89, 136, 156, 161, 316
Rodríguez, Tania, 285-286
Roose, Kevin, 117-118, 262
Rose, Charlie, 49
Rosenblatt, Frank, 96, 100-101, 103
Rosewood Hotel, jantar no, 37, 54-56, 62
Ruanda, 108
Rússia, 152
  guerra da Ucrânia, 59, 195

## S

Safe Superintelligence, 394
Salinas, Alejandra, 287
Sama AI, 194-196, 209-215, 220, 242, 403-404
Santiago, Chile, 269-271, 282, 284-285, 287, 292, 294-295
saúde e medicina, 24, 30, 83-84, 119, 229, 256, 300, 331
SB 1047, projeto de lei, 306
Scale AI, 198
  anotação de dados, 200-209
  projetos de RLHF, 215-219
  sistemas de pagamento, 208
Scallion, 365, 369-371
Schmidt, Florian Alexander, 199-200
Schulman, John, 218-219, 246, 257, 377, 393-394
  InstructGPT, 216-217, 219, 246
Schumer, Chuck, 51, 75, 307, 406
Scoble, Robert, 42
Scott, Kevin, 16, 75, 77, 79, 186-187, 247, 265, 268, 338
Sears, Mark, 209, 214
Sedol, Lee, 66
Segunda Guerra Mundial, 38, 97
segurança. *ver* IA, segurança de
segurança de IA. *ver* IA, segurança de
Seligman, Nicole, 366
*SemiAnalysis*, 266, 282
"senhas emergenciais", 154
Sequoia Capital, 41
servidores. *ver também* data centers
  Microsoft, 154
  OpenAI, 259-260
  refrigeração, 272, 275, 286, 288
sexismo, 120, 170, 232, 338
Shear, Emmett, 21, 43, 358, 361
Shopify, 54
Sidor, Szymon, 18-19, 150, 154, 244, 314, 358
Sierra, 366
Silverman, Carolyn, 29

simbolistas (simbolismo), 100-105, 114-115
Simo, Fidji, 366
"sinalização custosa", 350
síndrome de Ehlers-Danlos, 256, 331
"sinta a IAG", 126, 254
sistema de transporte rodoviário (NHTSA), EUA, 113
sistemas especialistas, 100, 102
sistemas fechados, 300, 304-305
Sky, 381-383
Slack, 11, 15, 21, 88, 116, 151, 161-162, 240, 261-262, 314, 350, 365, 379, 392
Slowe, Chris, 43
"Sobre os perigos dos papagaios aleatórios" (Bender), 170-179
sociedade limitada (LP), 73, 75-77, 90
Solon, Olivia, 108
Song, Dawn, 113, 119
Sora, 365
SpaceX, 33, 35, 37, 58, 360
Stable Diffusion, 120, 143, 236, 241, 281
Stack Overflow, 188
*StarCraft II*, 73
Starlink, 160
Steyerl, Hito, 143
Strawberry, 365-366, 394
Stripe, 49, 54, 62, 65, 80, 88-89, 243
Strubell, Emma, 165-166, 170, 176-178, 179,273-274, 305
Suleyman, Mustafa, 315, 375
Summers, Lawrence "Larry", 22, 366
SummerSafe, 75
Superalignment, 311, 346, 377-378.
Superassistant, 246-249, 256-257, 372
superinteligência, 29, 34, 60, 63, 96, 232, 310, 396
*Superinteligência* (Bostrom), 36-37
Suri, Siddharth, 196-97
Sutskever, Ilya
  "sinta a IAG", 126, 254
  Altman e, 341-343, 376-377, 387, 390-391
    comportamento de liderança, 339-352, 353-356, 373-376
    demissão e reintegração, 13-23, 355-363
  aprendizagem profunda e redes neurais, 106, 114
  conselho administrativo e de supervisão, 316-317
  cultura da OpenAI, 61-62
  desligamento de, 376-377, 390-391
  DNNresearch, 55, 58, 104-105
  entrevista da autora, 82-86
  expansão, 23-131, 165-166, 364-365
  estrutura de governança da OpenAI, 68-70
  fundação da OpenAI, 23-24, 36-38, 54-62

geração de código, 157
Hinton e, 55, 103, 105-106, 114, 116, 123, 127, 254, 258
histórico de, 54
ImageNet, 55, 67, 106, 103, 106, 123, 258
liderança de, 61-78
Murati e, 337-343
no Google, 54, 105-106
Omnicrisis e, 390-392, 394
paranoia de, 154
personalidade de, 15-16, 125-126
Projeto Manhattan de Alinhamento, 311
Q* (Strawberry), 364-365
Safe Superintelligence, 394
Superalignment, 311, 346, 377-378
Toner e, 320, 337, 342-352
Transformers, 126-127
Swift, Taylor, 14

## T

Tay, 159
Taylor, Bret, 22, 366
Te Hiku Media, 398-401, 405-406
*te reo Māori*, 397-400
Techworker Community Africa (TCA), 403
tecnonacionalismo, 304
Telêmaco, 277
Tenaya Lodge, 254
terra saqueada. *ver* extrativismo
"território de dados fluido", 295
Tesla, 70-72, 76, 93, 113, 115, 198, 338, 354
Model X, 76, 338
Murati na, 76, 338
piloto automático, 71, 113
Test of Time, prêmio, 258, 364
teste de estresse, 184
testes de QI, 97
testes padronizados, 97
texto-para-imagem, modelo, 181, 234, *ver também* DALL-E
Thiel, Peter
Altman e, 24, 35, 45-50
estratégia de "monopólio" de, 48, 147
Founders Fund, 47
fundação da OpenAI, 24, 48
Palantir, 38, 71
PayPal, 46, 76
Trump e, 46, 50
Threads, 259

Three Mile Island, reabertura, 273
TikTok, 299
Tiku, Nitasha, 253
*Time* (revista), 143, 195, 212, 404
Tironi Rodó, Martín, 271, 293, 296
Toner, Helen, 65
  Altman e o conselho, 13-14, 240, 317-318, 357
  comportamento de liderança, 339-352, 353-356, 373-376
  artigo sobre "sinalização custosa", 350
  Murati e, 337, 342-344, 348-350
  Sutskever e, 320, 337, 342-352
Tools for Humanity, 189-190
TPUs (unidades de processamento tensorial), 176
transcrição, 221
Transformers, 126-127, 129, 138, 234-235, 302
Transformers Generativos Pré-treinados. *ver GPTs específicos*
transparência, 17, 21, 25, 30, 57, 88-89, 92, 139, 149, 171, 177-178, 230, 297, 375, 393, 396, 407
transtorno de personalidade borderline, 331
triagem de resiliência, 210
Trump, Donald, 46, 50, 59, 199, 316, 395, 409
Tuna, Cari, 230
Turing, Alan, 88, 95, 97, 99, 111, 167, 364
Twitch, 21, 43, 358

## U

Uber, 111, 113, 116, 118, 142, 198, 228, 324
Ucrânia, guerra da, 59, 195
União Africana de Moderadores de Conteúdo, 403
União Europeia, 280
  EU AI Act, 306
*United Slate, The* (Altman), 50
Universidade Carnegie Mellon, 103, 111, 177
Universidade da Califórnia, Berkeley, 56, 63, 113-114, 124, 218, 235, 317, 406
Universidade da Califórnia, Los Angeles, 168
Universidade da Califórnia, San Diego, 103
Universidade de Chicago, 270
Universidade de Massachusetts Amherst, 86, 165
Universidade de Nova York, 111, 235
Universidade de Oxford, 63, 109, 229
Universidade Stanford, 34, 41, 48, 60, 81, 107, 110, 132, 138, 142, 147, 166, 235, 405
  AI Salon, 34
  Altman na, 41, 48
  Índice de IA, 110
Universidade de Toronto, 54, 111, 123
Universidade Yale, 199, 288
University of Applied Sciences Dresden, 199
Upwork Research Institute, 29

Uruguai, 270, 404
crise hídrica, 288-292
data centers, 288-295
uso de energia, 84, 86-87, 272-275, 277-279, 288-289
uso justo, 96
Utawala, Quênia, 194, 211, 213

## V

Vallejos, Rodrigo, 293-295
valores esperados, 229
"vantagem de sermos os primeiros", 109
Venezuela, anotação de dados, 198-209, 219
Venezuela, crise da, 198-201
véu da ignorância, 14
Villagra, Julia, 380
Vincent, James, 125
violação de direitos autorais, 96, 140, 297, 303, 308, 374
Virgínia, data centers, 276
vírus biológicos, 36
Volpi, Mike, 206
Volta, Alessandro, 138

## W

*Wall Street Journal*, 11, 42, 44, 49, 76, 108, 192, 196, 214, 278, 358, 374, 381, 403, 414
*WALL-E* (filme), 234
Wang, Alexandr, 205, 215, 219
Warzel, Charlie, 361
*Washington Post*, 75, 120, 253, 362, 390, 393
Watson Health, 105
Waymo, 106
Weil, Elizabeth, 320, 322, 326, 329, 332-333, 337
Weil, Kevin, 393
Weinstein, Emily, 303-304
Weizenbaum, Joseph, 101-102, 408
Welinder, Peter, 156, 160-161, 250
Weng, Lilian, 266, 395
West, Kanye, 222
West, Sarah Myers, 304
Whale, 277
Whisper, 244, 246, 265, 400
Wikipédia, 64, 131, 140-141, 222
Willner, Dave, 238, 249-251, 265, 395
WilmerHale, 366
Wong, Hannah, 255, 320, 322, 332, 414
Worldcoin, 190

## X

X (anteriormente Twitter), 15, 36, 141, 169, 173, 176, 219, 222, 237, 307, 312, 322, 329, 345, 366

xAI, 316-317, 387, 392, 394
Xerox PARC, 62

## Y

Y Combinator (YC), 33, 37, 41, 43, 45-52, 58, 316
Yom Kippur, 320
YouTube, 43, 59-60, 108, 141, 243-244, 327, 360
Yudkowsky, Eliezer, 314

## Z

Zaremba, Wojciech, 66, 157, 186
Zenefits, 45
Zilis, Shivon, 70, 315, 319, 375
Zoloft, 323-324
Zoph, Barret, 246, 372, 377, 394-395
Zuboff, Shoshana, 107
Zuckerberg, Mark, 47, 50, 165, 307, 316, 396

Título original
EMPIRE OF AI
Dreams and Nightmares in
Sam Altman's Open AI

Primeira publicação nos Estados Unidos por Penguin Press, um selo da Penguin Random House LLC, 2025.
Publicado com novo prefácio na Penguin Books, 2026.

*Copyright* © 2025, 2026 *by* Karen Hao

Todos os direitos reservados.

Nenhuma parte desta obra pode ser reproduzida, armazenada em sistema, ou transmitida, sob qualquer forma ou meio, sem a autorização prévia, por escrito, do editor, não podendo, de outro modo, circular em qualquer formato de impressão ou capa diferente daquela que foi publicada, tampouco pode ser utilizada ou reproduzida de qualquer forma para o treinamento, desenvolvimento ou operação de tecnologias ou sistemas de inteligência artificial (IA), incluindo tecnologias de IA generativa; sem as condições similares, que, inclusive, deverão ser impostas ao comprador subsequente.

Direitos para a língua portuguesa reservados
com exclusividade para o Brasil à
EDITORA ROCCO LTDA.
Rua Evaristo da Veiga, 65 – 11º andar
Passeio Corporate – Torre 1
20031-040 – Rio de Janeiro – RJ
Tel.: (21) 3525-2000 – Fax: (21) 3525-2001
rocco@rocco.com.br | www.rocco.com.br

Preparação de originais
ISIS BATISTA

Coordenação digital
MARIANA MELLO E SOUZA

Edição digital: maio, 2026.

CIP-Brasil. Catalogação na Publicação.
Sindicato Nacional dos Editores de Livros, RJ

H225i
Hao, Karen
O império da IA [recurso eletrônico] : por dentro da corrida irresponsável pela dominação total / Karen Hao ; tradução Ananda Alves, André Sequeira. - 1. ed. - Rio de Janeiro : Rocco Digital, 2026.
recurso digital

Tradução de: Empire of AI dreams and nightmares in Sam Altman's Open AI
Formato: epub
Requisitos do sistema: adobe digital editions
Modo de acesso: world wide web
ISBN 978-65-5595-484-5 (recurso eletrônico)

1. Inteligência artificial. 2. Livros eletrônicos. I. Alves, Ananda. II. Sequeira, André. III. Título.

26-104228.0 CDD: 006.3

CDU: 004.8

Carla Rosa Martins Gonçalves - Bibliotecária - CRB-7/4782
26/03/2026 27/03/2026

O texto deste livro obedece às normas do Acordo Ortográfico da Língua Portuguesa.

# A AUTORA

Karen Hao é formada em engenharia mecânica pelo MIT e jornalista premiada que investiga os impactos da inteligência artificial na sociedade. Ela escreve para publicações como *The Atlantic* e lidera a série AI Spotlight do Pulitzer Center, um programa que treina milhares de jornalistas mundialmente para cobrir os avanços da IA. Hao também foi repórter do *Wall Street Journal* e editora sênior de IA da *MIT Technology Review.* Seu trabalho é utilizado com frequência por universidades e governos ao redor do mundo. Ela já recebeu inúmeros prêmios, incluindo o American Humanist Media Award e o prêmio NEXT da American Society of Magazine Editors para jornalistas com menos de 30 anos.

# Índice de navegação


Capa
Folha de rosto
Dedicatória
Epígrafe
Sumário
Nota da autora
Prólogo
Parte I
Capítulo 1
Capítulo 2
Capítulo 3
Capítulo 4
Capítulo 5
Parte II
Capítulo 6
Capítulo 7
Capítulo 8
Capítulo 9
Parte III
Capítulo 10
Capítulo 11
Capítulo 12
Capítulo 13
Capítulo 14
Parte IV
Capítulo 15
Capítulo 16
Capítulo 17
Capítulo 18
Epílogo
Posfácio
Agradecimentos
Notas
Índice
Créditos
A Autora

# A adaga mágica

Gomes, Helena
9788581225265
42

Compre agora e leia (Publicidade)

Depois de passar seis anos trancafiado no mundo dos elfos, Razeel ganha o melhor presente de aniversário de Lyriel: a liberdade, mesmo que temporária. Porém, sua aventura se torna bem mais complexa e reveladora do que ele podia imaginar. Ele é levado ao passado para salvar a vida de Drake, mas acaba diante de uma antiga paixão que poderá tirá-lo novamente dos eixos. *A adaga mágica,* sequência de *O primeiro guerreiro,* é uma história da série **A Caverna de Cristais,** uma das sagas de fantasia/ficção científica pioneiras dentro da promissora literatura fantástica brasileira. Helena Gomes já tem mais de trinta obras publicadas, algumas delas selecionadas para

programas de leitura como o PNBE, e recebeu distinções importantes, como o selo Altamente Recomendável da FNLIJ.

Compre agora e leia (Publicidade)

# A herança da bruxa

Gomes, Helena
9788581224022
20

Compre agora e leia (Publicidade)

Décadas antes do nascimento do arqueiro Thomas, a jovem e inexperiente bruxa Couchet luta para sobreviver em Britanya, um reino dominado pelo preconceito e pela ignorância. Com a ajuda do elfo Razeel, a última bruxa da Ilha Média terá de se tornar a mestra de um poderoso guerreiro, peça-chave na luta desesperada que os aguarda no futuro contra os temidos nergals. *A herança da bruxa* é uma história da série **A Caverna de Cristais,** uma das sagas de fantasia/

ficção científica pioneiras dentro da promissora literatura fantástica brasileira. Helena Gomes já tem mais de vinte obras publicadas, algumas delas selecionadas para programas de leitura como o PNBE, e recebeu distinções importantes, como o selo Altamente Recomendável da FNLIJ.

Compre agora e leia (Publicidade)

# A hora da estrela

Lispector, Clarice
9786555950236
88

Compre agora e leia (Publicidade)

Pouco antes de morrer, em 1977, Clarice Lispector decide se afastar da inflexão intimista que caracteriza sua escrita para desafiar a

realidade. O resultado desse salto na extroversão é *A hora da estrela,* o livro mais surpreendente que escreveu. Se desde *Perto do coração selvagem,* seu romance de estreia, Clarice estava de corpo inteiro, todo o tempo, no centro de seus relatos, agora a cena é ocupada por personagens que em nada se parecem com ela.

A nordestina Macabéa, a protagonista de *A hora da estrela,* é uma mulher miserável, que mal tem consciência de existir. Depois de perder seu único elo com o mundo, uma velha tia, ela viaja para o Rio, onde aluga um quarto, se emprega como datilógrafa e gasta suas horas ouvindo a Rádio Relógio. Apaixona-se, então, por Olímpico de Jesus, um metalúrgico nordestino, que logo a trai com uma colega de trabalho. Desesperada, Macabéa consulta uma cartomante, que lhe prevê um futuro luminoso, bem diferente do que a espera.

Clarice cria até um falso autor para seu livro, o narrador Rodrigo S.M., mas nem assim consegue se esconder. O desejo de desaparecimento, que a morte real logo depois consolidaria, se frustra.

Entre a realidade e o delírio, buscando social enquanto sua alma a engolfava, Clarice escreveu um livro singular. *A hora da estrela* é um romance sobre o desamparo a que, apesar do consolo da linguagem, todos estamos entregues.
— **JOSÉ CASTELLO**, *Jornalista, escritor e Mestre em Comunicação pela UFRJ*

*"Esfinge, feiticeira, monstro sagrado. O renascimento da fascinante Clarice Lispector tem sido um dos verdadeiros eventos literários do século 21. Ninguém soa como Clarice. Ninguém pensa como ela. Ela não apenas parece dotada de mais sentidos do que os cinco conhecidos, mas também curva a sintaxe e a pontuação de acordo com sua vontade. Ela vira o dicionário de cabeça para baixo, soltando todas as palavras de suas definições, espalhando-as de volta como quer e não é que a língua parece melhor?"* — **THE NEW YORK TIMES**

Compre agora e leia (Publicidade)

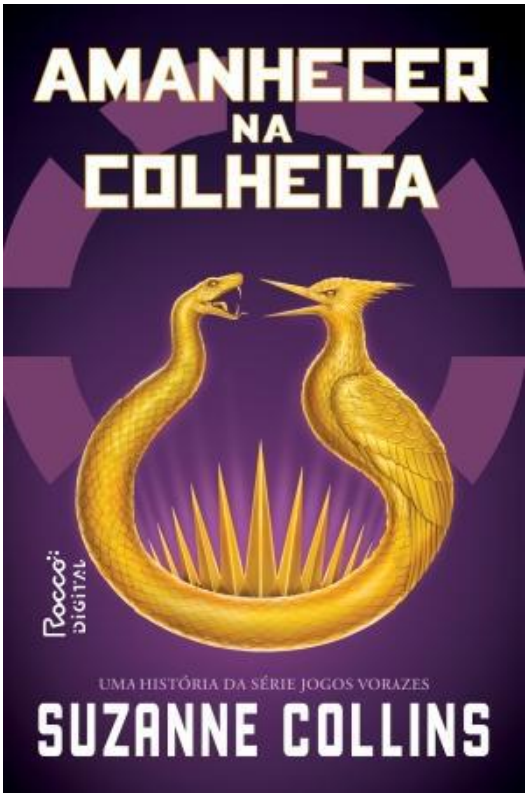
# O primeiro guerreiro

Gomes, Helena
9788581224671
36

Compre agora e leia (Publicidade)

Conheça a história de Kirian, o primeiro dos guerreiros eloras que renasceram no reino medieval de Britanya. Filho de Couchet, última bruxa da Ilha Média, e tendo como mentor o misterioso elfo, Razeel, o rapaz terá seu primeiro e definitivo confronto com os temidos nergals. *O primeiro guerreiro,* sequência de *A herança da bruxa,* é uma história da série **A Caverna de Cristais,** uma das sagas de fantasia/ficção científica pioneiras dentro da promissora literatura fantástica brasileira. Helena Gomes já tem mais de vinte obras publicadas, algumas delas selecionadas para programas de leitura como o PNBE, e recebeu distinções importantes, como o selo Altamente Recomendável da FNLIJ.

Compre agora e leia (Publicidade)




# Amanhecer na colheita

Collins, Suzanne
9786555953282
448

Compre agora e leia (Publicidade)

**Quando você está fadado a perder tudo que ama, pelo que ainda vale a pena lutar?**

Ao amanhecer do dia da colheita da Quinquagésima Edição dos Jogos Vorazes, o medo toma conta dos distritos de Panem. Nesse ano, em comemoração ao Massacre Quaternário, o dobro de tributos será levado de suas casas.

No Distrito 12, Haymitch Abernathy está tentando não pensar muito

nas suas chances de ser sorteado – só quer sobreviver ao dia e passar um tempo com a garota que ama.

Mas, ao ser escolhido, todos os sonhos de Haymitch desmoronam. Ele é separado da família e da namorada e enviado para a Capital com outros três tributos do Distrito 12: uma menina que considera quase uma irmã, um rapaz viciado em calcular chances e apostas, e a garota mais arrogante da cidade.

Conforme os Jogos se aproximam, Haymitch compreende que está tudo armado para o seu fracasso, mas parte dele deseja lutar... e deseja também que essa luta reverbere muito além da arena mortal.

Compre agora e leia (Publicidade)

---

O IMPÉRIO DA IA
Por dentro da corrida irresponsável pela dominação total
KAREN HAO
Rocco DIGITAL